Computational study of the metabolic diversity of the bacterium *Escherichia coli*

From single cells to cell communities and efficient systems

Eleftheria Tzamali

A thesis submitted for the degree of Doctor of Philosophy

Thesis Advisors: Dr. Martin Reczko, Dr. Panayiota Poirazi and Prof. Ioannis G. Tollis

Computer Science department
University of Crete
Greece

December 2010

UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE

Computational study of the metabolic diversity of the bacterium *Escherichia coli*

Dissertation submitted by
Eleftheria Tzamali

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Author:

Eleftheria Tzamali

Examination Committee:

Ioannis G. Tollis, Professor, University of Crete

Grigoris Antoniou, Professor, University of Crete

Panayiota Poirazi, Principal Researcher, IMBB-FORTH

Anastassios Economou, Associate Professor, University of Crete

Konstantinos Tokatlidis, Associate Professor, University of Crete

Konstantinos Marias, Principal Researcher, ICS-FORTH

Panagiotis Tsakalides, Professor, University of Crete

Approved by:

Angelos Bilas, Associate Professor
Chairman of the Graduate Studies Committee

Heraklion, December 2010

Ευχαριστίες

Πρώτα απ' όλα, οφείλω ένα πολύ μεγάλο ευχαριστώ στον καθ. Κ. Φωτάκη, διευθυντή του ινστιτούτου ηλεκτρονικής δομής και laser που με εμπιστεύτηκε, με ενθάρρυνε, με καθοδήγησε και με στήριξε στις όποιες ερευνητικές μου ανησυχίες και ερευνητικά βήματα από φοιτήτρια στο τμήμα Φυσικής μέχρι και σήμερα. Το ινστιτούτο αυτό αποτέλεσε έναν χώρο πάντα φιλόξενο για μένα και εκεί γνώρισα ερευνητές-δασκάλους-ανθρώπους μοναδικούς όπως ο Σ. Μαϊλης, ο Ν. Βάινος, η Β. Τορνάρη, ο Χ. Γρίβας και ο Δ. Άγγλος που αποτέλεσαν πηγή έμπνευσης και σημεία αναφοράς σε όλη μου την μετέπειτα πορεία. Τους ευχαριστώ βαθύτατα.

Θα ήθελα να ευχαριστήσω τον επιβλέποντα ερευνητή της διδακτορικής μου διατριβής, Martin Reczko, μια ήσυχη και βαθύτατα ειλικρινής παρουσία. Ο Martin με δέχτηκε με χαρά και με εμπιστεύτηκε αμέσως παρά το διαφορετικό μου υπόβαθρο. Ο Martin είναι ο άνθρωπος που σε αφήνει μόνο σου να διερευνήσεις και να ανακαλύψεις το επιστημονικό πεδίο και τα ενδιαφέροντα σου σε αυτό, ενώ σε παροτρύνει να μη φοβηθείς να δοκιμάσεις και να πειραματιστείς με τις ιδέες σου. Ένα πολύ θερμό ευχαριστώ στην επικεφαλής του εργαστηρίου Υπολογιστικής Βιολογίας, Παναγιώτα Ποϊράζη για το ενδιαφέρον που επέδειξε για τη δουλειά μου, την θερμή υποστήριξη της και την ενθάρρυνση να αγωνιστώ δυναμικά στις όποιες αντιξοότητες. Η Γιώτα με έκανε αμέσως μέλος του εργαστηρίου της και μαζί με τον Martin μου εξασφάλισαν τη διδακτορική υποτροφία μου μέσω του ΠΕΝΕΔ 03ΕΔ842. Η Γιώτα μας έφερε σε επαφή με δύο εξαίρετους επιστήμονες τον Α. Οικονόμου και τον Κ. Τοκατλίδη, καθηγητές στο πανεπιστήμιο Κρήτης και ερευνητές του ινστιτούτου Μοριακής Βιολογίας όπου και συζητήσαμε επί των αποτελεσμάτων αλλά και των κατευθύνσεων αυτής της δουλειάς και για τους οποίους είμαι βαθύτατα ευγνώμον. Ευχαριστώ τον υπεύθυνο της διδακτορικής μου διατριβής, τον καθ. Ι. Τόλλη που με το κριτικό πνεύμα του με ώθησε να προσπαθήσω για το καλύτερο και με στήριξε οικονομικά κατά την τελευταία περίοδο της εκπόνησης αυτής της εργασίας.

Ευχαριστώ θερμά τον Βαγγέλη Σακκαλή που έχοντας υπόψη του την αγάπη μου για τη βιολογία μου επέστησε την προσοχή στη συγκεκριμένη ερευνητική περιοχή, με καθοδήγησε με τον δικό του τρόπο και με παρότρυνε να συνεχίσω δυναμικά στους στόχους μου. Ευχαριστώ ιδιαίτερα τον Κώστα Μαριά, μέλος της εξεταστικής επιτροπής μου, ο οποίος ήταν εκείνος που με παρότρυνε να κάνω αίτηση διδακτορικού στο τμήμα επιστήμης υπολογιστών. Μέσω των υποχρεώσεων του τμήματος είχα επιπλέον την ευκαιρία να γνωρίσω εξαίρετους καθηγητές όπως ο Γ. Γεωργακόπουλος και ο Ι. Τσαμαρδίνος και να αποκτήσω γνώσεις πολύ καθοριστικές για την πορεία της διατριβής μου.

Ευχαριστώ τον φίλο μου Ιάσονα για την απέραντη υπομονή και την αγάπη του αλλά και τις πάμπολλες κουβέντες μας πάνω σε ποικίλα ερευνητικά θέματα που αδιαμφισβήτητα διεύρυναν τους επιστημονικούς μου ορίζοντες και όξυναν το επιστημονικό μου ενδιαφέρον. Ευχαριστώ τους πολύ καλούς μου φίλους, τη Νανσούλα, την Ήλια, το Μπουχλάκι, τον Κώστα, τη Γιώτα, τη Ελευθερίτσα και τη Δέσποινα, τη Μαρία με τον Παναή, τη Ρόη και τον Gfrag που βρίσκονταν σταθερά δίπλα μου ανεξαρτήτως απόστασης ή συνθηκών. Δεν μπορώ να παραλείψω το ζωικό μου βασίλειο, τόσο το εντός σπιτιού όσο και το εκτός, με την εκπληκτική ικανότητα τους να με ηρεμούν, να με διασκεδάζουν και να με κάνουν να νιώθω παιδί.

Ευχαριστώ τα μέλη των εργαστηρίων Υπολογιστικής Βιολογίας και Βιοπληροφορικής, την Κική Σιδηροπούλου, τη Νάση Παπουτσή, την Ελευθερία Πισσαδάκη, το Βασίλη Τσιάρα, τον Αναστάση Ούλα, τη Μαρία Μανιουδάκη, την Κατερίνα Γκίρτζου, το Νέστορα Καραθανάση, τη Σοφία Τριαναταφύλλου, την Ελένη Χριστοδούλου, το Γιώργο Καστελλάκη, τη Χριστίνα Φαρμάκη, τη Μαρία Ψαρρού και τη Δάφνη Κρυονερίτη που με τις ερωτήσεις τους, τις συζητήσεις μας και την ηθική τους στήριξη έκαναν τη δουλειά αυτή πολύ καλύτερη απ' ότι θα ήταν χωρίς τη δυναμική παρουσία τους. Ευχαριστώ τα παιδιά από το εργαστήριο της Ρομποτικής, το Νίκο Κυριαζή, το Μάρκο Σιγάλα, το Γιάννη Αυγουλέα, τον Κώστα Τζεβανίδη,

το Θωμά Σαρμή και τον Κώστα Παπουτσάκη για τη θερμή υποστήριξη τους. Επίσης ευχαριστώ το Μιχάλη Φλουρή που αμέσως βοήθησε σε ότι υπολογιστική ανάγκη προέκυψε.

Είμαι βαθύτατα ευγνώμον στην οικογένεια μου, τον Κωνσταντίνο και τα αδέλφια μου Κάτια και Αγγέλα. Η θερμή υποστήριξη τους, του καθενός με το δικό του τρόπο, σήμαινε πάρα πολλά για μένα. Ιδιαίτερα ευγνώμον είμαι στη μητέρα μου Αριάνα, έναν άνθρωπο δυνατό, περήφανο και ευαίσθητο που αποτελεί πάντα καταλυτική παρουσία στη ζωή μου και στην οποία αφιερώνεται αυτή η διατριβή.

To my mother Ariana

Abstract

In cross-feeding interactions, different strains or microbial species exchange usable products arising from the metabolism of the primal nutritional source. Cross-feeding interactions have been observed in several ecosystems. Furthermore, long-term evolution experiments on the bacterium *Escherichia coli* growing in a simple, single limited resource have shown the emergence of several subtypes with different phenotypes in the population maintained by cross-feeding interactions. Polymorphism and metabolic interactions can play an important role in the evolution of populations as they dynamically shape the fitness landscape allowing new phenotypes to evolve. Cooperative strategies in the form of cross-feeding may lead a population to better adaptation and more efficient exploitation of a given environment.

The availability of high-throughput data allows the mapping of cellular metabolism into a genome-scale metabolic network, which considers the set of almost all biochemical transformations that take place within the cell. Thus far, in the metabolic simulations, which describe bacterial growth based on the genome-scale metabolic reconstructions, cells are genetically identical.

In an attempt to improve our understanding of the evolution of metabolic diversity in simple environments and the mechanisms supporting cooperative behaviors, this work goes a step further from single-cell models; it develops the first genome-scale metabolic model capable of simulating a competitive life within cell communities, where different individuals co-grow, sense, shape and respond to a common, dynamic environment. The model aims to reveal communities composed of self-centered strains that exhibit group benefit because of their capability to better utilize the available resources than single strains. As proved analytically in this work, competition for the primal source alone in a simple and spatially homogeneous environment cannot lead a heterogeneous population to group benefit, supporting the hypothesis that other sources of heterogeneity such as by-production might play a critical role in growth efficiency. In addition to the metabolic model, a graph representation (diversity graph) is developed in order to reflect the mapping from the genetic to the metabolic variability with respect to by-production. The graph allows the efficient determination of strain communities with the potential to differently shape the environment and develop cross-feeding interactions. Several graph-theoretic measures are applied in order to reveal biologically insightful properties, to characterize the diversity graphs and to allow the direct comparison of the overall metabolic behavior of different mutants with respect to by-production under different growth conditions. The bacterium *E. coli* is used as a case study. Metabolic gene knockouts generate the pool of mutants among which potential cross-feeding interactions are examined.

The graph analysis suggests that the two acting processes towards stabilizing either the monomorphic (i.e. populations with a single mutant) or the polymorphic state are antagonistic and that among all potentially interacting communities probably only those consisting of mutants that are specifically adapted to the given environment are likely to evolve. It is observed that the metabolic capabilities of the mutants with respect to by-production are highly redundant. This property allows the efficient identification of all the potential interacting communities represented as cliques in the graphs. The growths of these communities are simulated in several growth conditions utilizing the developed genome-scale multi-competitor metabolic model. The growth simulations show that metabolic interactions are indispensable within strain communities in order to perform efficiently under conditions of resource competition. Strain communities can be beneficial even if not all of their pair-wise relations correspond to cross-feeding, which demonstrates the importance of exploring group-wise metabolic variability. Furthermore, it is observed that in several efficient cases co-growth provides immediate benefits to the competitors by increasing their growth rate. The existence of interacting heterogeneous populations capable of better exploiting a given growth medium than monocultures indicates that in some growth conditions, the involved

metabolic pathways are coupled in a way that a single optimal mutant is incapable of fully utilizing the environment.

As complexity increases and as environments become more complex than the homogeneous medium of a single-limiting resource that was explored in this study, diversity might prove far more beneficial for the systems involved. The method presented in this work has many implications for research on the ecology of increasingly complex microbial communities in natural and engineered environments.

Περίληψη

Η επιβίωση του 'καλύτερου' δεν είναι το μόνο πιθανό αποτέλεσμα στη διαδικασία της εξέλιξης. Εξελικτικά πειράματα σε βακτήρια έχουν δείξει πως ο πληθυσμός παρότι αναπτύσσεται σε ένα απλό, ομοιογενές περιβάλλον γρήγορα γίνεται και παραμένει πολυμορφικός. Επιπλέον, υποστηρίζεται ότι αυτή η πολυμορφία βασίζεται στις αλληλεπιδράσεις μεταξύ των διαφορετικών πληθυσμών κατά τις οποίες ανταλλάσσονται χρήσιμα για την ανάπτυξη παράγωγα του μεταβολισμού (cross-feeding). Τόσο η ποικιλομορφία όσο και οι μεταβολικές αλληλεπιδράσεις που αναπτύσσονται παίζουν σημαντικό ρόλο στην εξέλιξη ενός πληθυσμού βακτηρίων καθώς δυναμικά διαμορφώνουν το περιβάλλον ανάπτυξης επιτρέποντας την εξέλιξη νέων φαινοτύπων. Επιπλέον συνεργατικές συμπεριφορές με τη μορφή μεταβολικών αλληλεπιδράσεων μπορεί να οδηγήσουν έναν πληθυσμό σε καλύτερη προσαρμογή και καλύτερη αξιοποίηση του συγκεκριμένου περιβάλλοντος. Τα βακτήρια εμπλέκονται σε ποικίλες διεργασίες πάνω στον πλανήτη. Η ευρεία ποικιλομορφία τους και οι αλληλεπιδράσεις τους είναι αντικείμενο έρευνας σε περιοχές όπως η εξελικτική βιολογία, η οικολογία αλλά και η ανθρώπινη υγεία.

Μέχρι στιγμής η ανάπτυξη των βακτηρίων σε επίπεδο όπου ο μεταβολισμός τους περιγράφεται με μεγάλη ακρίβεια (genome-scale) έχει μοντελοποιηθεί για πληθυσμούς όπου όλα τα κύτταρα είναι πανομοιότυπα. Προκειμένου να κατανοήσουμε τη μεταβολική ποικιλομορφία όπως αυτή εξελίσσεται σε ένα απλό περιβάλλον αλλά και τους ακριβείς μηχανισμούς που περιγράφουν τις μεταβολικές αλληλεπιδράσεις σε έναν ποικιλόμορφο πληθυσμό κατασκευάζουμε, σε αυτή την εργασία, ένα νέο μεταβολικό μοντέλο ικανό να περιγράψει ποικιλόμορφες κυτταρικές κοινωνίες καθώς αναπτύσσονται σε ένα κοινό περιβάλλον ανάπτυξης, ανταγωνίζονται για τα θρεπτικά συστατικά, διαμορφώνουν και αλληλεπιδρούν με το περιβάλλον. Δείχθηκε αναλυτικά ότι μια κοινωνία ως σύνολο δεν μπορεί να είναι πιο αποδοτική από την απόδοση του καθενός μέλους της μεμονωμένα, αν το απλό περιβάλλον δε γίνει πιο σύνθετο όπως για παράδειγμα μέσω παραγώγων του μεταβολισμού και των μεταβολικών αλληλεπιδράσεων. Επιπρόσθετα με το μεταβολικό μοντέλο, ποσοτικοποιούμε και αναπαριστούμε γραφικά τη μεταβολική ποικιλομορφία γενετικά τροποποιημένων κυττάρων (diversity graph). Ο γράφος επιτρέπει τον προσδιορισμό μεταβολικά ποικιλόμορφων κυτταρικών κοινωνιών αποτελούμενων από κύτταρα με τη δυνατότητα να διαμορφώσουν διαφορετικά το περιβάλλον και να αλληλεπιδράσουν μεταβολικά. Οι ποικιλόμορφες κοινωνίες αντιστοιχούν στις κλίκες του γράφου. Διάφορες γραφο-θεωρητικές μετρικές εφαρμόζονται προκειμένου να καταλάβουμε τις ιδιότητες αυτής της αναπαράστασης αλλά και να συγκρίνουμε συνολικά τις μεταβολικές δυνατότητες του συστήματος κάτω από διαφορετικές συνθήκες ανάπτυξης. Η ανάπτυξη των διαφορετικών κοινωνιών προσομοιώθηκε και μελετήθηκε σε διαφορετικές συνθήκες ανάπτυξης αξιοποιώντας το προτεινόμενο μεταβολικό μοντέλο. Το βακτήριο *E. coli* αποτέλεσε τον οργανισμό προς μελέτη σε αυτή την εργασία. Επίσης, οι γενετικά τροποποιημένοι πληθυσμοί προέκυψαν μέσω διαγραφής ενός κάθε φορά γονιδίου εμπλεκόμενου στο μεταβολισμό. Η ποικιλομορφία μελετήθηκε κάτω από διαφορετικές συνθήκες αποτελούμενες από μία πηγή τροφής, πεπερασμένης ποσότητας.

Παρατηρήθηκε ότι οι γενετικά τροποποιημένοι κυτταρικοί πληθυσμοί έχουν στην πλειοψηφία τους παρόμοιες μεταβολικές ιδιότητες αναφορικά με την δυνατότητα παραγωγής παραγώγων του μεταβολισμού τους. Η παρατήρηση αυτή επιτρέπει το γρήγορο προσδιορισμό όλων των πιθανών κλικών–κοινωνιών του κάθε γράφου που κατασκευάστηκε για τις διαφορετικές συνθήκες ανάπτυξης. Η γραφο-θεωρητική ανάλυση επίσης υποδεικνύει ότι οι κοινωνίες που μπορούν να εξελιχθούν είναι πιο πιθανό να αποτελούνται από πληθυσμούς ειδικά προσαρμοσμένους στο συγκεκριμένο περιβάλλον. Οι αναλύσεις των κοινωνιών όπως προσομοιώθηκαν με βάση το προτεινόμενο μεταβολικό μοντέλο έδειξαν ότι οι μεταβολικές αλληλεπιδράσεις αποτελούν την αναγκαία συνθήκη προκειμένου μια κοινωνία κάτω από συνθήκες ανταγωνισμού των διαθέσιμων πηγών τροφής, να αξιοποιεί το συγκεκριμένο περιβάλλον πιο αποτελεσματικά από ότι ο κάθε διαφορετικός πληθυσμός μόνος του. Η ύπαρξη

τέτοιων αποδοτικών κοινωνιών υποδεικνύει ότι τα μεταβολικά μονοπάτια είναι συνδεδεμένα κατά τέτοιον τρόπο που υπό συγκεκριμένες συνθήκες ανάπτυξης δεν επιτρέπουν την ύπαρξη ενός κατάλληλα προσαρμοσμένου πληθυσμού ικανού να αξιοποιεί βέλτιστα το περιβάλλον του. Οι κοινωνίες μπορεί να είναι αποδοτικές ακόμα και αν δεν αλληλεπιδρούν ανά δύο όλοι οι διαφορετικοί πληθυσμοί, το οποίο αποδεικνύει τη σημασία της διερεύνησης της διαφορετικότητας ανά ομάδες. Παρατηρήθηκαν επίσης περιπτώσεις όπου οι μεταβολικές αλληλεπιδράσεις είχαν άμεσες συνέπειες στο ρυθμό ανάπτυξης των πληθυσμών που συμμετείχαν σε αυτές.

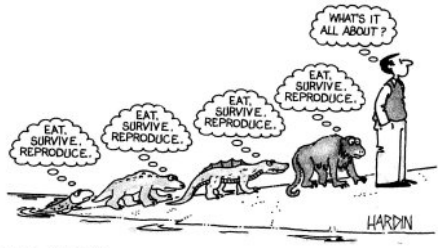
Καθώς η πολυπλοκότητα αυξάνεται και το περιβάλλον γίνεται πιο σύνθετο από το απλό περιβάλλον μίας πηγής τροφής, πεπερασμένης ποσότητας που μελετάται εδώ, είναι πιθανό η διαφορετικότητα να είναι πολύ πιο ευεργετική για τους πληθυσμούς που συμμετέχουν. Η μέθοδος που παρουσιάζεται σε αυτή την εργασία έχει εφαρμογές στη βιοτεχνολογία, στην ανθρώπινη υγεία για την κατασκευή αντιβιοτικών αλλά και στην οικολογία γενικότερα μικροβιακών πληθυσμών καθώς αναπτύσσονται είτε στο φυσικό τους περιβάλλον είτε σε τεχνητά περιβάλλοντα.

Contents

Glossary

Absolute benefit*	The relative difference between the performance of the community and the best single-growth performance among the pool of mutants (including the wild-type)
Biomass composition	The set of all precursors and building blocks that are essential for biomass production
By-product	Metabolite excreting during the metabolism of the primary nutritional source
Central node	A highly connected node
Clique	A fully connected sub-graph
Community*	A set of genetically and metabolically diverse strains
Competitive-exclusion principle	A single limiting resource can maintain only a single competitor
Cooperation	A behavior adopted by an individual (actor) that provides benefits to other individuals (recipients) and involves costs to the fitness of the actor
Cross-feeding	The exchange of by-products, which can be served as secondary resources in a population
Diversity graph*	A mapping from the genetic to the metabolic variability with respect to by-production. The Nodes correspond to the genetically different mutants. The Edge weights quantitatively describe the metabolic difference between two mutants with respect to by-production
Ecological niche	The set of environmental and ecological conditions under which a species persist. These conditions can be a nutritional source, physical conditions such as temperature or salinity
Efficient community*	A community that exhibits positive absolute benefit in a given growth condition and therefore performs better than any single-mutant from the pool of mutants
Evolutionary Retention Index (ERI) [39]	The fraction of genomes that have an ortholog of the given ORF with the number of representative organisms equal to 33. It takes values in [0, 1], where 0 corresponds to genes unique to *E. coli* and 1 for omnipresent genes
Fitness	A measure of the ability of a genotype to reproduce relative to other genotypes
Fitness landscape	The function that describes the relation between genotypes and fitness
Flux Balance Analysis	A method for predicting the flux distribution of chemical reactions based on constraints (i.e. stoichiometry, flux bounds) and optimization of an objective function (i.e. growth yield)
Frequency-dependent selection	The fitness of individuals depends on their frequency (relative abundance) in the population
Group*	A set of genetically and (optionally) metabolically diverse strains
Growth efficiency (Growth yield)	The ability of a system to maximize its growth performance (output) given a specific initial amount of resources (input)
Growth performance*	The maximum biomass concentration, which is the endpoint biomass concentration in metabolic dynamic simulations
Metabolism	The set of all biochemical reactions and transport mechanisms that take place in living organisms in order to maintain life
Multi-competitor growth model*	The proposed model for simulating the dynamic growth of heterogeneous cell populations in a common environment
Niche-exclusion principle	A single niche can support no more than one type whether it is a genotype or a species
Polymorphism	The existence of subtypes with different phenotypes
Relative benefit* Group benefit*	The relative difference between the performance of the community and the best single-growth performance among the members of the community
Stoichiometry	The quantitative relationships among substances as they participate in chemical reactions

***definition given in this work**



Nat. Hist., 1994, 103(6)

1. Introduction

1.1 Social life and Fitness

Mutation and selection comprise the core mechanisms for adaptation and evolution of biological populations. Fitness is the key criterion upon which natural selection acts and the ultimate arbiter of ecological success and survival. Fitness concerns the ability of a genotype to reproduce relative to other genotypes whereas it is always dependent upon the environment. The function that describes the relationship between genotypes and fitness is known as fitness landscape.

The evolutionary dynamic models usually describe the evolution of biological populations under the assumption that the environment does not change with the evolving population, which describes a constant, fixed fitness landscape. In a constant fitness landscape, the population evolves towards adaptations that maximize the fitness. However, it is possible that although the environment favors the adapted phenotypes, these adaptations shape the environment. An example is the selective advantage of a tree's height, which depends on the heights of the other neighboring trees or the selective advantage of a specific substrate, which when its concentration decreases an alternative substrate becomes more essential. A host's successful strategy to a pathogen may lead to newly adapted strains of pathogens and vice versa. A population evolving towards optimum adaptations changes the environment, which in turn shapes the optimum. Therefore, the fitness landscape can be dynamically shaped by the phenotypic distributions of the involved populations and their interactions, so that the success of a phenotype depends on the composition of the population [1, 2].

Darwin's theory of evolution states that individuals with traits that maximize their reproductive success will increase in frequency relative to other individuals. An interpretation of this theory is that natural selection favors species that boost their own success of offspring. Thus, in life individuals compete for survival and proliferation and eventually the fittest spreads within the population. However, across the animal kingdom, most communities are highly diverse whereas a variety of social behaviors have been observed. The African wild dogs, for example, have been observed to cooperate when hunting. The effectiveness of hunting depends on the number of cooperating hunters and the breeding success of the members of the pack also increases with group size. In several social vertebrates, coordinated guarding behaviors are observed where alternating individuals spend part of their time on guard protecting the group [3]. Cells cooperate in multi-cellular organisms. Cooperative metabolic activities have been also observed in microbial communities during degradation of organic pollutants [4]. Tumor progression is another example that has been proposed to be facilitated by cooperation among genetically diverse tumor cells through sharing of diffusible products [5]. Cooperation is evident at many levels of biological organization. Cooperative interactions are observed to most frequently evolve between relatives (kin selection), but even if rare are also developed in non-kin groups [6]. Darwin recognized that the evolution of cooperative societies posed a challenge to his theory of natural selection. Cooperation is a difficult behavior to explain. The problem is that individuals

who do not cooperate, but gain benefit from the cooperative behaviors of others, will gain a competitive edge and invade the population.

In a continuously changing, multitasking and complex real world, diversity within a population might provide the opportunity to better utilize the available resources, adapt, survive or perform by interacting with each other than when functioning as individuals. However, natural selection is not a far-sighted process and individual selection usually acts more rapidly than group selection. Thus, cooperative strategies are more likely to be established in natural communities when evolution favors their formation among self-centered individuals due to the emergence of immediate benefits at the individual level.

1.2 Bacterial Diversity

Diversity rapidly emerges even in the simplest environments and the simplest organisms and microbial life does not comprise an exception. Microorganisms or microbes are microscopic unicellular organisms that first appeared on earth 3-4 billion years ago. Microbes comprise the richest repertoire of molecular and chemical diversity, whereas a variety of social behaviors involving communication and cooperation have been observed [7-9] showing that microbes are social, interacting systems in the same manner as many other species on earth. Microbes keep the planet running. Nevertheless, the actual underlying mechanisms that drive natural systems to stable polymorphism and allow cooperative social behaviors in microorganisms to emerge still comprise an open issue with a long debate [7, 10-13].

Examples of microorganisms include bacteria, fungi, archaea and protists. Bacteria are single-celled, prokaryotic microorganisms, which are observed to frequently live in dense populations forming biofilms whereas they interact with each other through a variety of secreted molecules. Bacteria have certain appealing properties such as short generation times, large population sizes and ease to grow, that make them an excellent model to studying ecological and evolutionary processes. The broad diversity and social behaviors of microbes are thus an important subject in the area of evolution, social evolution and ecology. However, bacteria are also of great importance because they underlie many ecosystems and biotechnological processes. Understanding bacterial diversity -apart from its biological significance- is of great importance in more applied areas such as bio-degradation of pollutants, in food preservation as well as in human health. Bacteria are ubiquitous in every habitat on earth. Throughout life, humans become hosts of bacterial cells that are greater in number than the human cells. The evolution of bacterial pathogens and their interactions with each other and the environment are of particular importance since the extensive variability of pathogens within populations continues to threaten human life.

1.3 Bacterial population growth

1.3.1 Binary fission

Binary fission is the process that all prokaryotes, several protozoa and some eukaryotic organelles such as mitochondria and chloroplasts use for asexual reproduction and cell division.

Through binary fission a living prokaryotic cell is divided into two daughter cells. The process begins with DNA replication. Prokaryotes have a single circular chromosome. It is attached to the inside of the plasma membrane. The cell membrane

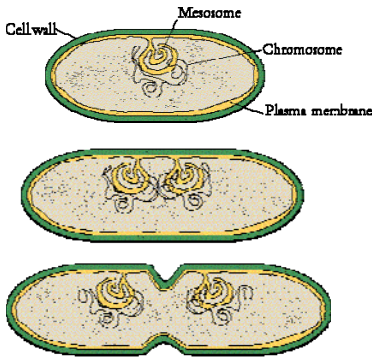


Figure 1.1. Prokaryotic fission
Source: http://home.earthlink.net/~dayvdanls/pcelldiv.htm

elongates and once the cell doubles its original size the cell membrane begins to invaginate (pinch inwards). A cell wall is formed between the two DNA copies and the cell is divided into two daughter cells, which are identical with each other if no mutational event has occurred during the process. Considering that on average the number of daughter cells surviving the fission exceeds unity then the bacterial population increases exponentially.

1.3.2 Laboratory growth environments

Batch culture is the most common laboratory growth environment, in which bacterial growth is studied. The bacterial culture is incubated in a closed vessel with a single batch of medium. Ideally the medium is well-stirred and thus spatially homogenous (unstructured), whereas the concentrations of the substrates vary over time characterizing the environment as temporally structured. Serial batch culture is when passage of the culture into fresh medium is conducted. Serial-transfer culture is analogous to a seasonal environment in which resources are in abundance at the beginning of a transfer cycle and then as the population approaches its saturation density they become scarce.

In other experimental regimes, some of the bacterial culture is periodically removed while a fresh sterile media is added. In the extreme case, the fresh medium of the nutrients is continuously added, while culture liquid is continuously removed, so that the culture volume remains constant. This is the case of a chemostat also known as continuous culture. It is ideally spatially unstructured and temporally unstructured. One of the most important features of chemostats is that micro-organisms can be grown at a physiological steady state. In steady state, growth occurs at a constant rate and all culture parameters remain constant (culture volume, dissolved oxygen concentration, nutrient and product concentrations, pH, cell density, etc.). In addition environmental conditions can be controlled by the experimenter.

Most laboratory techniques for growing bacteria use high levels of nutrients to produce large amounts of cells cheaply and quickly. However, in natural environments nutrients are limited, meaning that bacteria cannot continue to reproduce indefinitely. In nature, many microorganisms live in communities (e.g. biofilms) which may allow for increased supply of nutrients and protection from environmental stresses. The cross-feeding relationships that might be developed in a population can be essential for growth of a particular organism or group of organisms, a phenomenon called syntrophy.

1.3.3 Growth phases

Bacterial growth in a batch culture normally follows four phases (Fig. 1.2). The first phase of growth is the *lag phase*, a period of slow growth where bacteria are synthesizing, at high rates, RNA, enzymes and other molecules as a process of adaptation to growth conditions. During the lag phase bacteria are maturing and not able to divide. The second phase of growth is the *exponential phase,* which corresponds to the period of cell doubling. This phase is also known as *logarithmic (log) phase* because cellular growth is usually represented on a logarithmic scale. During this phase, nutrients are metabolized at maximum rate. The number of divisions per cell per unit time is called *growth rate* and corresponds to the slope of the line of the population growth represented on a logarithmic scale. The growth rate depends on the organism, the growth condition and the probability of the

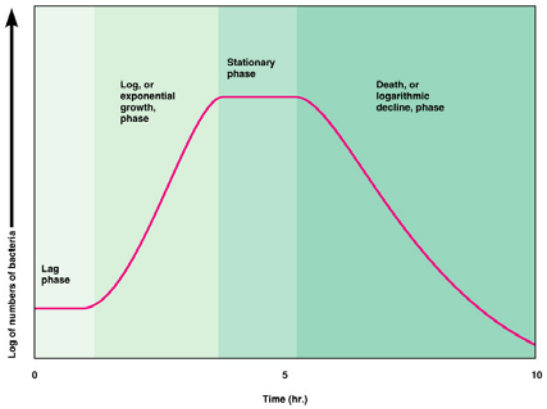


Figure 1.2. Bacterial growth phases.
Source: http://www.microscopesblog.com/2009/11/bacterial-growth.html

daughter cells to survive in each doubling. The time it takes the cells to double is known as *generation time*. As nutrients are depleting and the environment becomes enriched with wastes growth starts to slow down. In a chemostat culture bacteria are maintained in the exponential growth phase. The next phase is the *stationary phase*, which is characterized by a vanishing growth rate because of the nutrient depletion and accumulation of toxic products in the medium. The bacterial growth rate equals the bacterial death rate. The stationary phase is a transition from rapid growth to a stress response state and there is increased expression of genes involved in DNA repair, antioxidant metabolism and nutrient transport. The final phase is the *death phase*. Death is usually exponential.

1.3.4 Fitness and Relative fitness assays in bacteria

In the long-term evolutionary experimental studies [14] that will be presented later, fitness is usually assayed by the reproductive success over successive generations. In that respect, the fitness of a bacterial population is in direct relation with the number of doublings occurring within a specified time interval. Let $N(0)$ be the initial population and $N(t)$ be the bacterial population at time t, then the number of doublings at time t equals to $D(t) = \ln(N(t)/N(0))/\ln 2$.

On the other hand, relative fitness can be assayed by placing two different populations in competition. In this case, the fitness of the strain population i relative to the strain j is given by $W_{ij} = D_i(t)/D_j(t)$.

1.4 Metabolism

Metabolism includes the set of all biochemical reactions and transport mechanisms that take place in living organisms in order to maintain life. Metabolism is at the core of cellular function whereas it comprises one of the most fundamental and well-conserved cell processes of living organisms [15, 16]. The central pathways of metabolism such as the citric acid cycle and glycolysis are found in all three domains of life (archaea, bacteria and eukaryotes). Through metabolism, the energy sources that are available in the environment are converted into ATP and other cellular building blocks to construct essential components such as proteins and nucleic acids in order for the cell to maintain growth and survival (Fig. 1.3). The energy sources differ from organism to organism. Metabolism can be divided into two processes; catabolism and anabolism. Catabolism concerns the set of metabolic reactions that break down large molecules to provide energy. Anabolism then utilizes the energy to synthesize essential components such as proteins and nucleic acids. Metabolic imbalance is related to many human diseases such as diabetes, obesity, cancer, cardiovascular disease and Alzheimer's disease.

Most of these chemical reactions are thermodynamically unfavorable and for that reason they require the presence of the appropriate enzymes to take place. Regulatory mechanisms interacting with the cell's internal and external environment are responsible to provide to the cell the proper enzymes at the proper times making metabolism subject to regulation. Metabolism bridges the cell's environment with the cell's phenotype and reflects interactions from signal transduction to genetic regulation and protein level making its study essential for understanding how cells function.

Metabolic networks are flow networks. The metabolic network represents the channels for the flow of material and generation of Gibbs free energy, which are constrained by the conservation laws of mass and energy. Beside the internal set of fluxes, fluxes entering and leaving the boundary of the cell (exchange fluxes) through the transport mechanisms are present as well. The flux represents the flow of material through a reaction channel. It

depends on the amount of available reactants, the amount of enzyme that catalyzes the specific reaction, the affinity between enzyme and substrate. As enzymes are subject to regulatory mechanisms, so is the flux of the corresponding biochemical reaction. The enzyme activity is programmed by the organism to provide certain chemical compounds at certain times in certain environmental conditions so that the organism accomplishes its metabolic demands as they emerge.

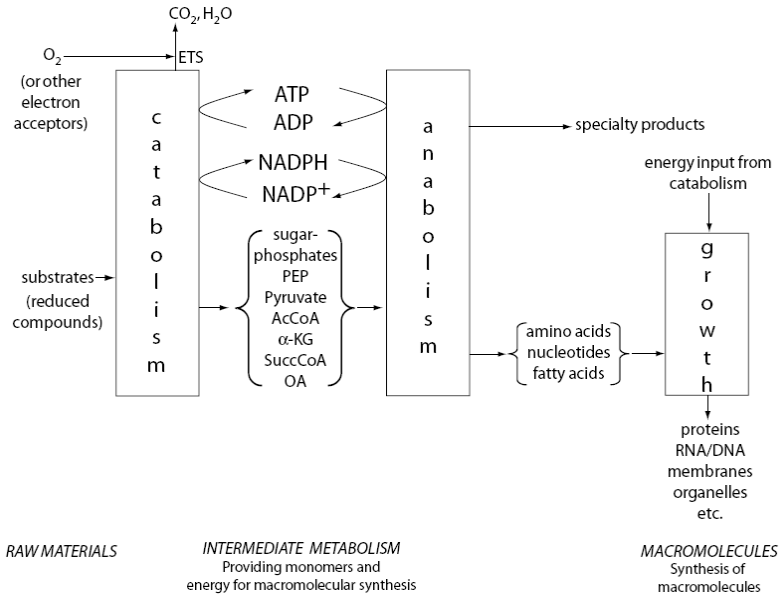


Figure 1.3. A coarse-grained representation of cellular metabolism presented in [17].

The metabolic network reconstruction as well as the state of the art in the metabolic modeling is analytically described in Chapter 3.

1.5 The competitive-exclusion Principle

The niche-exclusion principle [18] states that no more than one type whether it be a genotype or a species can occupy the same niche for long time. A niche corresponds to the set of environmental and ecological conditions under which a type persists. When the environmental condition corresponds to a single limiting resource the principle is known as competitive-exclusion principle. In a complex growth environment where a variety of available niches such as different resources or physical conditions are possible, different types may be favored in each niche. Diversity can thus be maintained according to the niche-exclusion principle. However, bacterial species co-inhabiting an ecological niche have also been observed. The evolution of non-transient polymorphisms in asexual populations growing in simple environments seems to contradict to the competitive-exclusion principle that supports that after successive clonal replacements a single genotype, the 'fittest' will dominate the population. It is, however, possible that the growth of one genotype creates a metabolic opportunity (new niche) for another genotype, as it dynamically shapes the environment giving rise to metabolic interactions between the members of the population that will in turn stabilize the polymorphisms.

1.6 Long-term evolution experiments on *E. coli*

Long-term evolution experiments on monoclonal populations of the bacterium *Escherichia coli* growing in a homogeneous environment consisting of a single-limiting resource have verified

the emergence and maintenance of more than a single strain in the population under either chemostat or serial batch culture [19-22].

A. Chemostat culture experiments

A population of *E. coli,* initiated with a single clone and maintained in long term, *glucose*-limited, chemostat culture developed extensive polymorphism [19]. 773 generations of growth under highly selective conditions have failed to produce a single strain of superior metabolic capabilities; that is maximum *glucose* uptake and maximum ability to metabolize secondary metabolites. Three evolved strains, the CV103, CV116 and CV101, were isolated after the 773 generations and studied under competition experiments. The reconstruction experiments revealed that the evolved polymorphism involves cross-feeding of residual metabolites such as *acetate* and *glycerol* that were produced during the metabolism of *glucose* (Fig. 1.4). These metabolites comprised alternative resources for growth in the population transforming the simple, *glucose*-limited environment into a complex one that allowed the evolution of balanced polymorphism. The exact metabolic capabilities of the three persisting genotypes are the following. Genotype CV103 possessed superior *glucose* uptake kinetics among the three strains. This genotype was observed to exhibit low growth yield, a capability of producing *acetate* and *glycerol* but it was incapable of consuming *acetate*. Genotype CV116 showed a higher maximum growth rate and an enhanced ability to assimilate *glycerol*. Monocultures of CV116 exhibit equilibrium *acetate* levels lower than those of CV103 but significantly higher than those of the genotype CV101.

Furthermore, the evolved clones differed significantly from one another with respect to their metabolic capabilities and their global gene expression patterns. According to Kurlandzka et al [23] changes were observed in the expression levels of proteins associated with translation, membrane composition, shock response and active transport whereas a decreased number of synthesized proteins were observed in all adaptive clones.

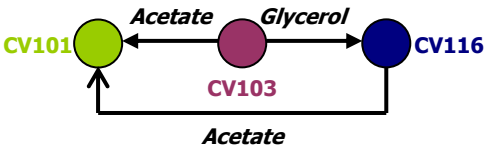


Figure 1.4 The metabolic interactions as described in [19] between the three evolves strains CV101, CV103 and CV116.

B. Serial transfer population experiments

The longest-running evolution experiment was initiated in 1988 by Lenski [24]. The experiment is an ongoing study that uses the bacterium *Escherichia coli* as experimental organism. Twelve independent populations were initiated from the same ancestor, a clone of *E. coli* B, and propagated by daily serial transfer in the same *glucose*-limited environment.

Genetic changes have been reported within the evolving populations and many studies analyzed the alterations that emerged. Parallel and convergent changes in gene and protein expression patterns were observed under the same culture conditions indicating adaptive diversification evolution [25, 26]. Most of the differences in gene and protein expression between the evolved and ancestral strains involved a down regulation in the evolved strains.

In addition, the long-term evolution experiments show that while the evolving populations adapted well to the *glucose* medium, unused catabolic functions decayed and their diet breadth became narrower and more specialized [27]. The experiments show that antagonistic pleiotropy, in which the same mutations are beneficial in one environment and detrimental to another, was the main contributor to the resource specialization during evolution of *E. coli* lines in minimal *glucose* medium. The decay of the unused catabolic functions is the result of

adaptive mutations that improve fitness on *glucose.* The mutation accumulation that occurs stochastically cannot explain the functional losses.

The initial 2000 generations correspond to the period of the most random adaptations in which the within population genetic variance of fitness is observed to correspond to the observed rate of adaptation as predicted by the Fisher's fundamental theorem [14]. Surprisingly, however, after 10000 generations the rate of genetic adaptation significantly declined whereas the variation in performance among clones within a population remained at high levels [28]. The emergence of two distinct genotypes, *L* and *S*, in a dynamic stability of their coexistence was identified in the evolving population at generation 18000. The two genotypes differed with respect to the average size of their individual cells. Furthermore, the maximum growth rate of type *L* was significantly higher than *S*. While both morphotypes secreted at least one metabolite into the medium, the cross-feeding interaction mainly benefited *S* by increasing its growth rate. The fitness of *S* relative to *L* in competition over the entire growth cycle was found to be (negative) frequency-dependent, indicating a selective advantage when the genotype is rare. The evolved polymorphism and strong frequency-dependent interactions were not observed in other replicate population growing in identical conditions. This indicates that the chronological order of the mutations that occur in an evolving population plays an important role in the evolution of polymorphism.

Summary of the experimental findings

1. After several generations of growth in a simple, *glucose*-limited environment, the population of *E. coli* initiated with a single clone shows stable polymorphism. The metabolites produced during the metabolism of *glucose* transform the simple environment into a complex one.
2. The polymorphism is maintained by cross-feeding (metabolic) interactions or/and demographic trades off (negative frequency-dependence) that protect rare genotypes from extinction.
3. All adaptive clones whether grown on batch or chemostat culture, exhibit a decreased number of proteins synthesized when compared to the parent strain. Furthermore, the global gene and protein expression patterns significantly differed from one another.
4. The evolved populations exhibit substantial difference in certain phenotypic traits, such as average cell size and performance in novel environments. Difference in the maximum growth rate was also observed between the evolved populations. The maximum growth rate is a very important fitness component and if all other features were identical coexistence wouldn't have been observed.
5. The chronological order of the mutations that occur in an evolving population plays an important role in the evolution (or not) of stable polymorphisms.
6. As the bacterial populations adapt to the *glucose* medium, loss of catabolic functions occurs. The adaptive mutations improve fitness on *glucose* while affecting the performance of the populations in other substrates. Consequently, the populations become more specialized exhibiting a narrower diet breadth.

1.7 Sources of phenotypic variation in Bacteria

Mutations are the main reason for the origin of genetic variation among individuals within bacterial populations. The evolutionary success of bacteria relies on the mutation rate, which can be shaped by many factors such as competition, environmental heterogeneity, and population size [29]. Beside point mutations, which are mainly responsible for the fine-tuning of gene functions, genetic and functional studies combined with data from comparative genomics have revealed the major importance of both gene acquisition and gene loss to the emergence and evolution of bacterial pathogens and symbionts [30]. In the absence of genetic mutations, stochasticity in gene expression, which is present due to finite number

effect, has been also proposed as a mechanism for generating phenotypically distinct subpopulations within isogenic cell populations as it may affect protein abundance [31, 32].

Genetic perturbations, such as the inactivation or activation of a gene involved in the metabolism, may exhibit no or only a mild effect on the phenotype. This is either because a single gene participates in a set of inactive metabolic reactions for a given environmental condition such that its variation has no effect on the metabolic capabilities of the cell or because of the evolved functional redundancy that characterizes many cellular systems allowing them to be robust in a variety of genetic perturbations and environmental conditions [33-38]. Genome-scale deletion phenotype data for the bacterium *E. coli* [39], for example, show that 83% of the protein-coding genes (~87% of the total were examined) were dispensable in rich growth media. Metabolic simulation data for the *E. coli* under 30000 diverse simulated environments show a relatively low number (11.9%) of metabolic reactions that were always active [36]. Nevertheless, specific genes might prove either indispensable for cell viability or their perturbations may considerably alter the metabolic capabilities of the cell. The evolution of a single strain of improved metabolic capabilities - the 'fittest' that eventually replaces all others is expected. On the other hand, specific viable mutants may exhibit specific metabolic properties, which might prove beneficial in an interacting population. Furthermore, natural selection may favor individuals of certain metabolic capabilities to coexist with other individuals in a population, if this leads the population to efficient growth, for example by metabolically interacting with each other.

1.8 Problem Statement

In cross-feeding interactions, different strains or microbial species exchange usable metabolic products arising from the metabolism of the primal nutritional resource. Cross-feeding has been observed in several ecosystems. Furthermore, polymorphism has been observed to emerge from a monoclonal bacterial population that grows in a single-limited resource and sustain in the population when beneficial cross-feeding interactions between the members of the population take place. Polymorphism and metabolic interactions can play an important role in the evolution of populations as they dynamically shape the fitness landscape allowing new phenotypes to evolve. Cooperative strategies in the form of cross-feeding may lead a population to better adaptation and more efficient exploitation of a given environment.

The understanding of the underlying mechanisms regarding the (bacterial) diversity and its maintenance and the exploration of novel, interesting phenotypes, which might be derived from diverse interacting populations entail the reconstruction of models capable of bridging this genotype-phenotype gap.

Although single-cells have been described at genome-scale, detailed genome-scale models capable of explaining the mechanisms involved in polymorphic populations and cooperative societies are missing. The state of the art presented in the following focuses on three research directions; the development of genome-scale models and their predictive capabilities, alternative approaches that have been suggested in order to analyze cross-feeding polymorphisms and theoretic models that analyze their evolutionary stability.

1.9 State of the Art

Cells are complex dynamical systems that are constantly remodeling themselves in response to changes in their internal and external environments. A lot of research effort has been devoted to the development of whole cell, *in-silico* genome-scale metabolic models [40-43] aiming to globally describe the complex interactions and processes that take place within a living cell. These metabolic models are based on a genome-scale metabolic network reconstruction, which includes the set of almost all biochemical reactions that take place within a cell. The reconstruction of well-curated metabolic models is in general a hard and

time consuming problem and these reconstructions are available for a small, but rapidly increasing number of organisms [43, 44]. The functional states of the reconstructed networks are usually explored under the constraint-based framework [40-43]. Flux Balance Analysis (FBA) models [41, 45] are widely used constraint-based models that utilize the genome-scale metabolic network and estimate the *optimal* flux distribution of the entire biochemical reacting system, providing a quantitative description of the system when the intracellular fluxes are in balance. Dynamic cellular growth has been also described under the FBA framework at the population level [46, 47]. The populations consist however of identical cells, which follow the same metabolic and regulatory program. The genome-scale metabolic reconstruction and modeling are described in detail in chapter 3 as they are an essential component of this study.

The constraint-based framework has been successfully used to analyze the metabolic capabilities of several organisms [48-50], among which the most represented domain is bacteria. Genotype-phenotype relations at genome-scale have been broadly studied under the FBA framework with the aim of identifying essential genes or metabolites, exploring the metabolic robustness, identifying minimal metabolic networks and investigating their evolution, inferring the lifestyle of an organism as well as comparing and validating the metabolic networks or even designing organisms with a desirable metabolic phenotype [43, 51-53]. In most previous studies, genetic perturbations are explored with respect to their effect on cellular growth yield, viability or productivity of specific metabolites that have an industrial significance. Nevertheless, the method also enables the prediction of the relative flux values of the metabolic reactions [54]. In fact, the ability of the model to reliably predict the intracellular and transport fluxes under several genetic perturbation and growth conditions is subject to the way evolution has shaped the operational criteria that lead the organism to survival and growth [55, 56]. Anyhow, besides the growth phenotype predictions, the model can also provide information about the metabolic by-product secretion of the cell, which has shown to be consistent with experimental data for specific environmental conditions [47, 57]. This thesis utilizes this information and examines computationally the effect of the specific genetic perturbations on the metabolic capabilities of the derived mutants with respect to by-production.

Based on the existing *in-silico* models, several research studies in the field of metabolic engineering area have focused on the identification of genetic manipulations that can result in efficient microbial strains with desirable, improved phenotypes such as the growth yield [58-61]. These efforts however have focused on the search of the fittest, monoclonal population. The current work examines whether metabolically interacting strain communities, as opposed to monoclonal populations, can provide an alternative way of identifying growth efficient systems.

Very little work has been done on the interrogation of multi-species interactions and the analysis of polymorphisms at the genome-scale. The work of Stoylar et al. [62] is the first reported reconstruction of a dual-species stoichiometric model, which focuses on the central metabolism rather than the genome-scale and which applies a common objective function (reconstructing a bi-level optimization) in order to analyze the syntrophic association between the microbes *Desulfovibrio vulgaris* and *Methanococcus maripaludis* in several environments. The underlying hypothesis here is that under several evolutionary pressures and long periods of coexistence, loss of individuality may occur resulting in changes in the initial strategies of each individual and eventually its replacement by a common objective.

Without detailed biochemical description of the cells, several parametric models have been proposed in order to determine the conditions that allow the evolution of stable coexistence of competing species. In these models, the species are attributed with certain growth properties and metabolic capabilities according to the experimental observations while most models rely on growth-associated product formation kinetics. Several microorganisms that compete for a limiting nutritional resource and exchange intermediate products of metabolism

have been modeled [62-65]. Despite their simplicity, the models produce predictions that are qualitatively verified by the experiments.

To avoid the complexity of the variables that can play an important role in the establishment of cooperation in natural systems, Shou et al [65] constructed a simplified synthetic obligatory cooperative system for studying the evolution of cooperation. The system was composed of a pair of yeast strains, each of which produced a nutrient required by the other. The viability of the system relies on cross-feeding and its stability over a wide range of conditions was shown both mathematically and experimentally.

The evolution of cooperative behaviors and stable polymorphisms in biological systems has been also investigated theoretically under the well-established game theoretic framework [2, 63, 64, 66-68]. Game theoretical models such as the Prisoner's dilemma for pairwise interactions and the public good games for groups of interacting individuals have been widely applied to biological problems. A *game* represents an interaction among individuals in which *players* act according to their phenotypes, known as *strategies*. As biological systems evolve in a dynamic fitness landscape, the success of a strategy depends on the phenotypic distribution of the population. The gradual evolution of cross-feeding polymorphism from a single ancestral microbial strain has been nicely analyzed, under the evolutionary game-theoretic approach [69]. In the model of Doebeli [69], the maximal growth rates of the primal (*glucose*) and secondary (*acetate*) resources are the two phenotypic traits that are allowed to evolve. The model predicts the specific conditions for the emergence of cross-feeding in both chemostat and serial batch cultures. The model also suggests that polymorphism is less likely to evolve in serial batch cultures because very high rates of production of the secondary nutrient are required.

1.10 Thesis overview

This thesis aims to develop a predictive mathematical and computational framework capable of exploring cross-feeding polymorphisms, identifying interacting bacterial communities and simulating the growth of heterogeneous cellular population at genome scale. A description of the interacting population at genome-scale allows a detailed understanding of the metabolites that are exchanged and of the way the growth rates and other metabolic capabilities can be shaped under conditions of resource competition and cross-feeding. Contrary to existing simplified metabolic models, the interacting populations are allowed to dynamically shape, adapt and utilize the given environment according to their metabolic capabilities. Under the plausible hypothesis that the best use of resources enhances the likelihood of survival subject to ecological and evolutionary constraints, communities built out of self-centered strains of efficient, improved growth performances, capable of better utilizing the available resources comprise an important aspect of this study and are thoroughly investigated.

This goal of this thesis is not to study evolutionary time scales or prove long-term stability of the predicted polymorphisms; it describes however the co-growth of interacting 'selfish' strains as dynamic adaptations to a shaped environment, an approach which is consistent with the evolutionary optimization theory [1, 2].

An analytic description of the proposed method is presented in chapter 4.

1.10.1 The search space

The current work selects the bacterium *Escherichia coli,* a well-studied and best characterized organism in terms of its genome annotation and functional characterization, as a case study. Genetic perturbations and specifically single, metabolic gene knockouts generate the pool of mutants among which potential cross-feeding interactions are examined. As previously seen

(section 1.6), gene inactivation or loss comprises a common strategy of the adaptation process of a bacterial population to a specific, laboratory environment.

Within a specific pool of mutants this work explores computationally the ecologically relevant, metabolic diversity that emerges and may evolve from a monomorphic state in a bacterial system, which grows in a simple, unstructured environment consisting of a single-limited resource.

An exhaustive search of all the possible compositions of mutants is impractical because of the exponential search space. Within a specific pool of n mutants, the size of the search space for a given community size k is equal to the binomial coefficient $\binom{n}{k}$. However, the compositions of metabolically interacting strain communities are searched among individuals with different metabolic capabilities under the assumption that metabolically diverse strains have the potential to differently shape the given environment and provide each other with useful products of their metabolism. This assumption can significantly reduce the search space.

When the organism is shifted from one growth condition to another, the active metabolic pathways adapt either through changes in the fluxes of already active reactions (flux plasticity) or by changing the active wiring system (structural plasticity) [36]. Therefore, strains exhibit different metabolic capabilities in different sources. The emerged metabolic variability, the compositions of the potential metabolically interacting communities as well as their growth are thus source-dependent.

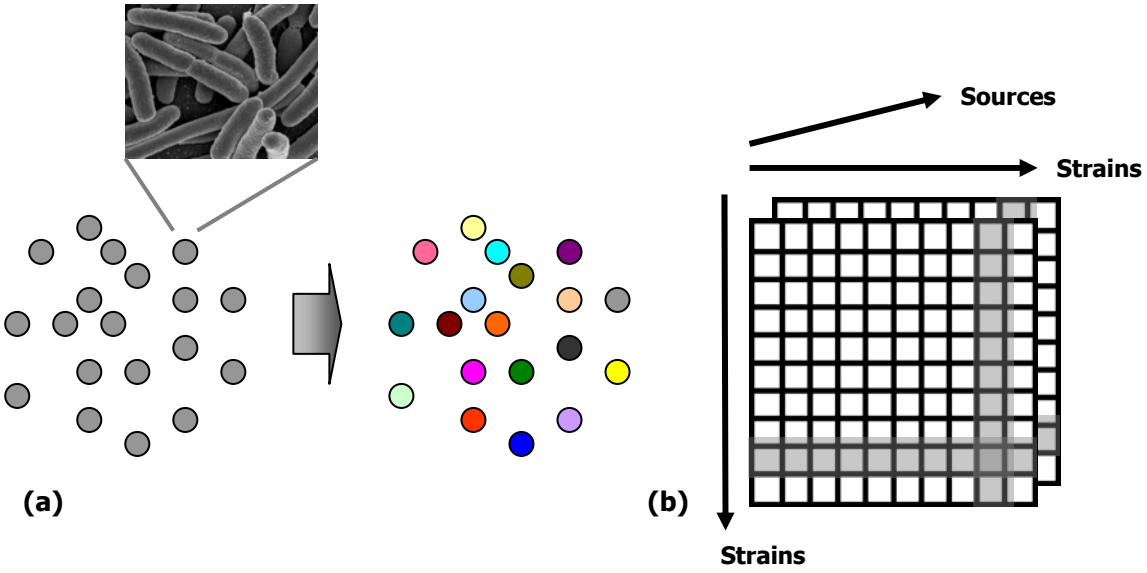


Figure 1.5 The search space for exploration of single mutants and mutant pairs. (a) Genetically different cells (depicted with color) derived from single-gene knockouts of the wild-type cell of *E. coli* (depicted with grey color) produce the strains that are examined in this work for their potential to develop cross-feeding interactions. (b) The strain-strain potential interaction space for different growth conditions (sources). A sub-space such as the wild-type – strain interactions is highlighted.

1.10.2 Investigated systemic properties

- What is the maximal metabolic variability with respect to by-production that can emerge within a pool of genetically different cells under a given growth condition? How complex a simple, single-source environment can be?

- Which are the growth conditions where the organism is observed to be more sensitive-fragile to genetic perturbations increasing its probability to develop polymorphism?

- Which strains have the potential to develop cross-feeding interactions? Which are the potential metabolically interacting strain communities that can emerge from a pool of strains under a given condition?

- How evolutionary stable a strain with high potential to develop cross-feeding interactions is?

- Are there specific strains appearing as consistent members of the potential interacting communities across the diverse growth conditions?

- How does a diverse strain community behave metabolically under conditions of resource competition, considering that each individual functions towards its best survival and proliferation?

- Are there communities capable of better utilizing the available resources than the monoclonal populations? How important are the cross-feeding interactions?

- Does the community consist of more diet-specialized strains?

1.10.3 The proposed methodology

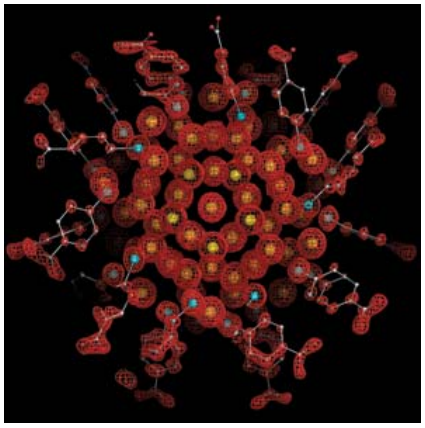
This work develops, for the first time, a genome-scale metabolic model capable of simulating the co-growth of any number of different cells in a common, dynamic medium where each different cell type functions towards its best survival and proliferation. The new algorithm is called multi-competitor metabolic model and is based on the existing genome-scale metabolic model [47], which describes growth of monocultures. Inputs of the model include the genome-scale metabolic networks of the individuals, their constraints and the (initial) growth conditions. Output of the model is the time-varying optimum flux distribution, which determines temporarily the concentrations of the resources and the population growth of each individual in the heterogeneous population.

The pool of the genetically different mutants is systematically investigated with respect to the metabolic capabilities of the mutants regarding the by-production. The metabolic products disturb the homogeneity of the growth environment, comprise the mean of exchange in an interacting population and provide the raw material of interesting phenotypes to emerge. The metabolic capabilities of the strains are determined based on the single-cell *in silico* metabolic model. A graph representation is developed in order to map the genetic to the metabolic variability and to allow the efficient determination of the metabolically different pairs of strains as well as the identification of all the potential bacterial communities that consist of strains, which under the same initial conditions have different metabolic blueprints. The graph provides a non-centralized, comparable, quantifiable and extendable description of the metabolic diversity whereas it enables the identification of the potential metabolically interacting individuals. Diversity graphs can be reconstructed in order to identify the potentiality to exchange products, ideas, or information within a group depending on the discipline. Several graph measures are applied in order to reveal biologically insightful properties, to characterize the diversity graphs and to allow the direct comparison of the overall metabolic behavior of this organism with respect to by-production under different growth conditions when single-gene knockouts are applied.

Given the novel, multi-competitor growth model, many interesting and diverse directions that can be explored arise such as the final population frequency, frequency-dependent relationships, and (initial) resource amount-dependencies on the cross-feeding interactions. This study focuses on whether cross-feeding interactions can lead a community to efficient growth.

1.11 Thesis Organization

Basic concepts of network analysis with emphasis on methods applied on biological systems are described in chapter 2. Basic concepts of systems biology and specifically on the genome-scale metabolic network reconstruction and the constraint-based framework are presented in chapter 3. The proposed methodology of this work, which is mainly divided into the introduction of the multi-competitor algorithm and the reconstruction of a graph to describe potential cross-feeding interactions are described in chapter 4. Chapter 5 analyzes the results of this study. An exhaustive analysis of different wild-type-mutant growth simulations is first performed under 58 different carbon sources using the proposed multi-competitor algorithm. The results of this analysis are presented in section 5.2. The metabolic simulations reveal specific conditions where specific wild-type-mutant pairs are capable of utilizing the given environment more efficiently than single-cell populations due to cross-feeding interactions. The model identifies the exact metabolites that are exchanged in these interactions and the way cross-feeding shapes the metabolic capabilities of the involved cells. All the beneficial wild-type-mutant pairs show high metabolic variability with respect to by-production between the involved mutant and the wild-type. This analysis subsequently inspired the diversity graph reconstructions. The diversity graphs are reconstructed for each growth condition. We study these graphs in section 5.2 using graph-theoretic measures presented in chapter 3. The graph analysis shows that the metabolic capabilities of the mutants with respect to by-production are highly redundant. Environmentally invariant mutants, which consistently appear as either highly different (central) or highly redundant (non-central) in the diversity graphs, are observed. It is also observed that most environmental-invariant, highly central mutants correspond to deletions of highly conserved genes. The results of the simulations of all the cliques-communities found in the diversity graphs of *glycolate*, *acetate*, *glycine*, *glucose, pyruvate* and *melibiose* are presented in section 5.3. The growth simulations show that strain communities can be beneficial even if not all of their pair-wise relations correspond to cross-feeding, which demonstrates the importance of exploring group-wise metabolic variability. Furthermore, it is observed that beneficial metabolic interactions can be either bi-directional where the exchange of essential nutrients flows in both directions, so that both mutants exploit the newly shaped environment or they can be in one direction where only one benefits from the coexistence and the other plays the role of a mere provider, an altruist. It is observed that as long as pairs of mutants of unexplored or unexploited metabolic capabilities are not present, the performance of the clique linearly depends on the mean performance of the pair-wise interactions. In the Appendix of this thesis it is shown analytically how the mass conservation does not allow group benefit to emerge under conditions of competition of the primal nutritional resource. Chapter 6 summarizes the main conclusions and presents future directions and possible applications of the proposed framework.



2. Basics in Network Analysis

2.1 Introduction

Network or graph analysis has been widely used to describe systems ranging from social interactions, power grids and Internet to neural connections in brain, protein interactions, ecosystems, evolutionary relationships and evolutionary games [70-75]. Network-based representations of biological systems can provide insights into their underlying machinery as well as their structural and functional organization and their evolutionary origin [74, 76]. A network representation also allows comparative analysis [77] to be performed as well as visualization tools [78, 79] to be applied on biological systems. Network analysis attempts to give answers to important questions such as how information, diseases or pathogens spread in a population, how molecules interact with each other and the environment allowing cell functioning and more general how local interactions determine global behaviors of networks. Statistical physics has played an important role in the development of graph theoretic tools for the understanding of complex systems regarding its ability to connect the microscopic dynamical evolution of the components of the complex systems with the emergent microscopic phenomena. Algorithms borrowed from graph theory as well as tools and properties proposed for the analysis of complex networks can be applied in biological systems if they are described as networks.

This section includes a general introduction to graphs attempting to present networks within a biological context. The most basic graph vocabulary and conventions that will be used throughout this study are presented. In-depth description of graph theory is beyond the scope of this study. An interested reader is referred to [80, 81] for a deeper understanding of graph theory. An introduction to the analysis of networks and the reconstruction of network models is also presented. Extensive analysis of complex networks with a focus on either their structure or the physical and dynamical processes and interactions occurring among their constituent elements can be found in [76, 82].

2.2 Network definition

A network or graph is a representation of the interactions among the components of a system. Examples of network components include genes, proteins, metabolites, species, people, routers and html documents. In biological systems, the interactions can be physical, representing the potential of physical contact between two components (such as protein-DNA binding) or temporal, describing the influence a component will have on another component in a posterior time, thus tracking the temporal transition of each component of the system, or they may depict dependencies among components under certain environmental conditions. Any kind of meaningful integration of the above is also possible.

The components of interest are depicted as nodes in the network whereas the interactions between these components are depicted as edges (see examples in 2.2.1). Thus, in a formal definition, a graph $G = (V, E)$ consists of a set of vertices-nodes V and a set of edges $E \subset VxV$.

Each component can also be seen as a variable, whose values describe the component's population levels. Each edge is defined by the two nodes it connects. These nodes are called the end-points of the edge. An edge can either be directed (arc), representing the flow of information, material, dependence or causality or it can be undirected representing simply the interaction between the components. A graph augmented in a way that represents directed edges is called *directed graph* or *digraph*.

A *subgraph* of a graph G is a graph whose node set is a subset of the set of nodes of G and whose edge set is a subset of the edge set of G. The "removal" of a node from G must be followed by a "removal" of all the edges that are incident with the node.

The *complement* or *inverse* of a graph G is a graph H on the same set of nodes such that two nodes in H are adjacent if and only if they are not adjacent in G. A graph G whose edges reflect a kind of similarity has a complement graph H whose edges represent dissimilarity.

A *simple graph* is an undirected graph, which does not contain loops (edges started and terminated at the same node) or multiple edges between the same pair of nodes.

2.2.1 Weighted Networks

Edges can also carry weights, which depending on the problem can either reflect cost, capacity, reaction speed, and distance, degree of the interdependence or confidence level between two components of the network. Graphs with value-assigned edges are called *weighted graphs*.

A weighted graph $G = (V, E, w)$ is a graph with a function $w : E \to R$ that assigns a real number to every edge of the graph.

A *grayscale* network consists of a set of vertices, a set of edges and a function, which assigns a real number in (0, 1] to every edge and corresponds to the strength of the dependence between vertices. The value 0 corresponds to the absence of an edge. Self-loops are not allowed. Grayscale networks were introduced in [78] in order to describe and visualize brain activity. They are called gray simply because their edge values can be visualized using different shades of gray. These networks can be considered a subclass of the weighted networks. The measures and the analysis of these networks can be applied in any network where the weight values are naturally constrained in (0, 1].

2.2.2 Examples of biological networks

Cells are complex biological systems that consist of components such as genes, gene products and metabolites that interact with each other selectively and dynamically at different levels in the cascade from genes to proteins and metabolites in response to internal and environmental signals. A traditional approach to study graphically the inherent complexity of cellular or other real systems is by decomposing the whole at different levels and focusing on one type of component (e.g., proteins) and on a specific type of relationship between them.

At the genomic level, the transcription of genes to mRNAs is regulated by transcription factors, which are themselves products of genes. In gene regulatory networks, genes and gene products comprise the nodes of the networks whereas the edges represent the assumed influences between them. A simplified, small-sized gene regulatory network is shown in Figure 2.1.

At the post-translational level proteins participate in interactions that can lead to modified protein functions or the formation of various protein complexes. These processes are reflected in the protein-protein interaction networks. An example of the protein-protein interaction network of the yeast *Saccharomyces cerevisiae* as derived from yeast two-hybrid measurements [83] is presented in Figure 2.2.

Metabolism includes the set of all biochemical reactions that are essential for an organism to maintain life. Metabolism is commonly represented as a flow network where the nodes are metabolites and the edges establish the metabolic reactions that are catalyzed by certain enzymes (Figure 2.3). In an alternative representation, the nodes of the metabolic network correspond to metabolic genes whereas an edge among two components exists if and only if there is a metabolite that is catalyzed by enzymes encoded by both components.

Several different representations are usually possible for the same system of components, depending on the algorithms used and their underlying assumptions. As a result, the kind of influence we want to represent (physical, temporal, conditional), the underlying assumptions regarding the system behavior (deterministic, stochastic), the level of detail in the system's observations (from Boolean to continuous or stochastic variables) and the algorithms used, define the qualitative and quantitative information that can be incorporated into and/or inferred from a model.

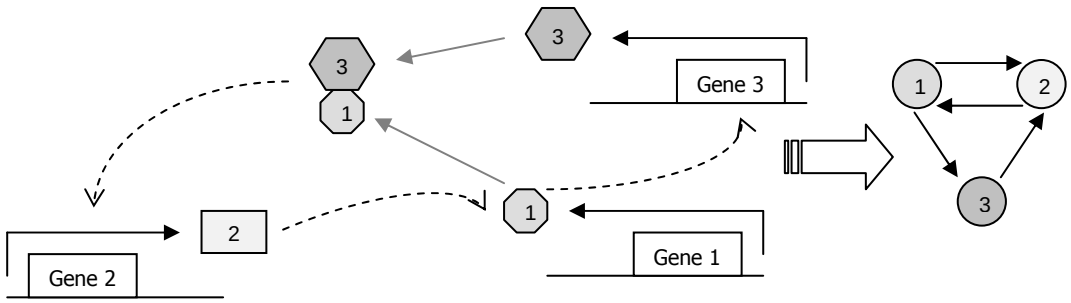


Figure 2.1 A small gene regulatory system and its corresponding network representation. Genes are translated into proteins which may form complexes as the gene products 1 and 3; these proteins or complexes regulate in turn the genes, thus forming a network of interactions. Discontinuous arrows indicate regulations.

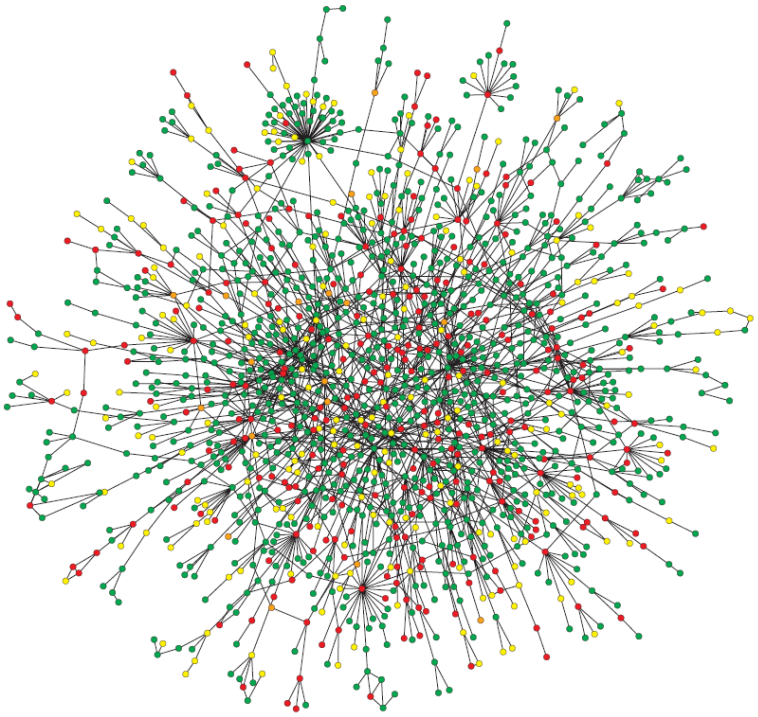


Figure 2.2 Map of protein-protein interactions in *Saccharomyces cerevisiae*. The network is reconstructed based on yeast two-hybrid measurements. The color of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). The figure is taken from [74].

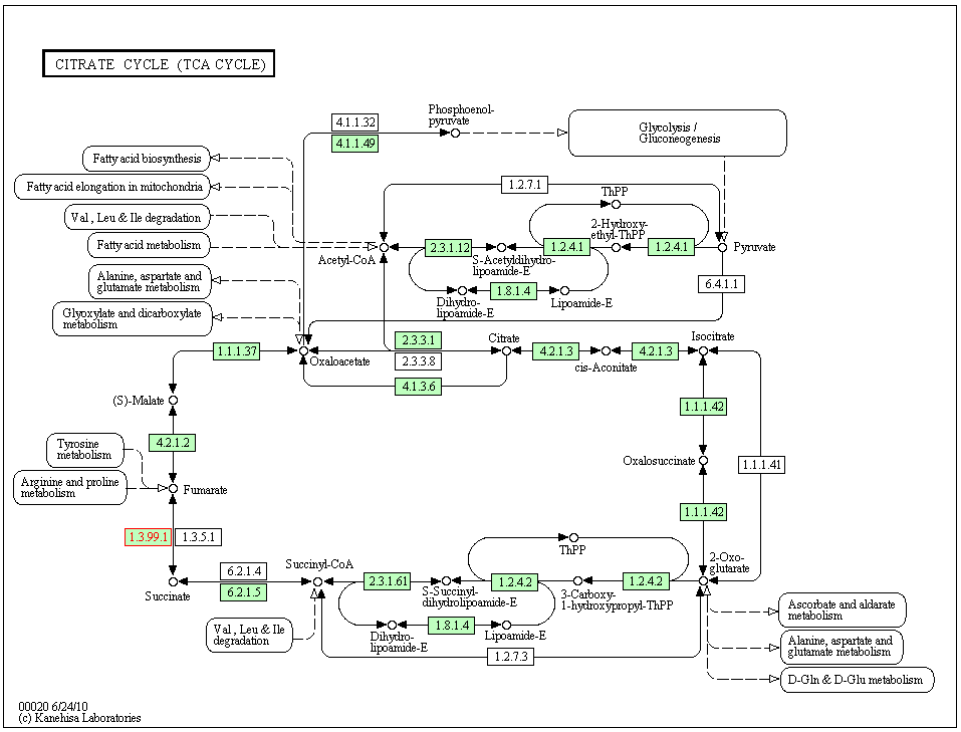


Figure 2.3 The citrate cycle (TCA cycle, Krebs cycle) of the bacterium *Escherichia coli* as presented in the Kegg database [84]. The TCA is an important aerobic pathway for the final steps of the oxidation of carbohydrates and fatty acids. The circles depict metabolic compounds whereas the genes involved in the metabolic pathways are shown between the compounds.

2.3 Network analysis

The understanding of the structural organization, function and evolution of networked systems ranging from the World Wide Web to social and biological systems entails systematic network-based analysis and tools capable to explore and characterize the inherent complexity of these networks. Centrality measures have been widely applied to characterize local properties of networks and to analyze the role the components of a network play in the functioning of the system. Originally the measures were defined for binary representations of the corresponding graphs, in which edges are either present or absent. Nevertheless, edges may have weights encoding valuable information about the relations of the network. In an attempt to utilize this information, several studies extend the network-based measures to weighted graphs [78, 85, 86].

2.3.1 Graph Density

The maximum number of edges in a simple graph of n nodes is equal to $\frac{n(n-1)}{2}$ and corresponds to a *complete* graph. All the nodes of a complete graph are connected with each other. A *dense* graph is a graph with number of edges close to the maximal number of edges. If m is the number of edges of a simple, undirected graph of n nodes then the graph density D is given by the number of existing edges m over the maximal possible number of edges. Thus, $D = \frac{2m}{n(n-1)}$. A graph is *sparse* if $D \ll 1$ whereas a complete graph has $D = 1$.

2.3.2 Node Centrality

2.3.2.1 Degree in binary graphs

The most fundamental feature of a node in a graph is its *degree*. The node degree is equivalent to the simplest centrality measure known as *degree centrality*. The degree of a node v, written $d(v)$, is defined as the number of edges incident with v. The degree is a local measure, which expresses the importance of a node in a graph with respect to its connections.

In a graph of n nodes the maximum possible degree of a node equals to $n-1$. The degree can be normalized to [0, 1] if divided by the maximum possible degree of the graph [87]. The normalized degree or centrality of a node v is thus:

$$c_D(v) = \frac{d(v)}{n-1}$$

A node with high centrality is also known as hub.

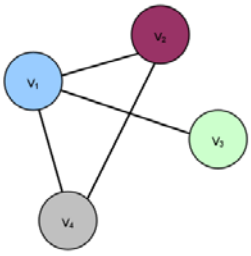


Figure 2.4 A simple graph of 4 nodes and 4 edges. The degree of each node is: $d(v_1) = 3, d(v_2) = 2, d(v_3) = 1, d(v_4) = 2$

The node centrality has been related to biological characteristics such as essentiality in biological networks. Deletion analyses, for example, have shown that the degree of a protein

in the protein network of *S. cerevisiae* has an important functional role determining its deletion phenotype. Many hub proteins have been shown to be essential for the survival of the cell [83]. Furthermore, comparative analysis of the metabolic network of 43 different organisms revealed the same set of highly connected nodes-metabolites indicating the generic utilization of the same substrates by each organism [88].

A graph $G=(V,E)$ of n nodes and m edges is called *b-regular* if $d(v)=b, \forall v \in V$. The total number of edges, m of this graph equals to $\frac{nb}{2}$. A complete graph is $(n-1)$-regular.

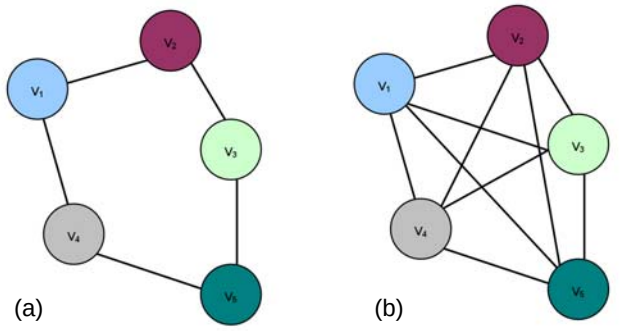


Figure 2.5 An example of a 2-regular graph (a) and a complete graph (b).

In directed graphs, the edge (v_i, v_j) is said to be *incident from* v_i and *incident to* v_j. For directed graphs two different degrees are defined; the *in-degree*, which corresponds to the number of the incoming edges and the *out-degree*, which corresponds to the number of the outgoing edges. Thus, for a node v the in-degree $d^-(v)$ represents the number of edges incident to v, whereas the out-degree $d^+(v)$ depicts the number of edges incident from v.

A directed graph is *balanced* if $d^+(v)=d^-(v), \forall v \in V$.

2.3.2.2 Strength in weighted graphs

The degree has been extended to take into account the weight values of the edges when analyzing weighted graphs [85, 86]. The *strength* of a node also known as *strength centrality* is determined by the sum of the weights of the edges incident with the node. The normalized strength of a node v, $C_s(v)$ corresponds to the mean value of the edge weights w of its neighbors. This definition is also applicable to the binary graph representations presented previously where edges take values in $\{0,1\}$. In that case, the node strength corresponds to the node degree. In grayscale networks [78], the normalized strength centrality takes values within [0, 1].

$$C_s(v)=\frac{1}{n-1}\sum_{e=\{v,u\}\in E} w(e)$$

2.3.3 Distance and paths

A *path* from a node v_i to a node v_j in a graph is an ordered sequence of distinct edges and nodes, linking a source node v_i to the target node v_j. A *cycle* is a closed path in which the source and the target nodes are the same.

The *length of a path* equals to the number of the traversed edges. In weighted graphs where the weights of the edges represent distance, the length or *weight of the path* corresponds to the sum of the weights of the traversed edges.

The *shortest path* from a source to a target node corresponds to the path of minimum length. In network analysis, it is important to calculate the shortest path of all pairs of nodes in the graph. Several algorithms exist that solve this optimization problem efficiently, the most common of which are the Floyd-Warshall algorithms with a runtime that grows asymptotically as fast as V^3 and the Johnson's algorithm that grows asymptotically no faster than $V^2 \log V + VE$.

The maximum shortest path of a graph is called *graph diameter.* A graph diameter is the longest path in the graph when paths which backtrack, detour or construct circles are excluded from consideration. In a comparative analysis of the metabolic networks of 43 species representing all three domains of life, Jeong et. al. [88] observed that all networks had the same diameter. This finding was not expected and a possible explanation would be that a larger diameter attenuates the organism's ability to adapt efficiently to environmental changes and internal errors where the synthesis of more enzymes would have been required. The *mean path length* represents the mean shortest path of a graph and provides a measure of a network's overall navigability. A short mean path length implies that local perturbations in the nodes of the graph could reach the whole network fast; that is they display enhanced signal-propagation speed. This effect is known as *'small-world effect'* and has been observed in several biological and technological networks.

2.3.3.1 Distance in weighted graphs

The weight values assigned on the edges of the weighted graphs might not correspond to distance length between two nodes and they might not directly be interpreted as lengths. If for example, the edge values reflect any kind of dependence, relatedness, similarity, potential to interact then the weight values should at least reverse in order to reflect distance.

To express distance between the edges of this kind of weighted graphs is important when measures that involve shortest paths are to be explored. An example of a monotonic non-increasing function g that can be applied on the weights, w_{ij} of the edges is:

$$g(w_{ij}) = 1 - \log_2(w_{ij})$$

Under that definition, edges with weight value equal to zero have infinite distance, whereas edge weights equal to 1 have distances equal to 1, expressing direct links. Other transformation are given in [78].

2.3.4 Cliques

Cliques in undirected graphs are complete subgraphs; that is, they consist of a subset of the nodes of the graph in which all the nodes are connected with each other. The size of a clique corresponds to the number of its nodes. The *maximum clique* is a clique of the largest possible size in a given graph.

A clique of size k is comprised of sub-cliques of size $m < k$ the number of which correspond to the binomial coefficients $\binom{k}{m}$. A *maximal clique* is a clique, which cannot be extended by

including one more adjacent node from the node set of the graph. A maximal clique is also called inclusion-maximal because it does not exist exclusively within the node set of a larger clique.

In a social network, the maximum clique can represent the set of individuals in which all know each other. In protein-protein interaction networks [89] and in gene regulatory networks [90] groups of highly densely interconnected nodes also known as modules can provide important information regarding the organization and functioning of the cellular system.

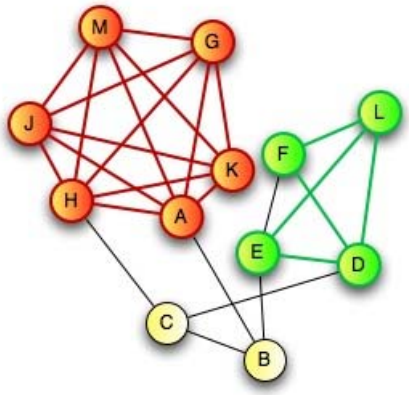


Figure 2.6 Maximal cliques are depicted with different node colors. The maximum clique of the graph is the red clique consisting of the six nodes: $\{A, K, G, M, J, H\}$.

The identification of the maximum clique size in a graph is an *NP-complete* problem, which practically means that for arbitrary graphs there is no polynomial time algorithm known. The clique problem and the *independent set* problem are complementary. An independent set is a set of nodes in a graph not two of which are adjacent. A clique in a graph G corresponds to an independent set in the complement graph of the graph G. As a result many computational algorithms are applied equally well to either problem. However, for a restricted family of graphs the two problems might not be equivalent as for example for *planar* graphs. In that specific case, the clique problem can be solved in polynomial time [91] whereas the independent set problem remains NP-hard for planar graphs.

Many algorithms have been developed [89, 92, 93] for the identification of cliques or independent sets in graphs since the problem apart from its theoretic interest has many practical applications. The theory has been devoted to identifying special families of graphs that allow efficient algorithms capable of solving the problem fast. On the other hand, approximation algorithms have also been developed.

2.3.5 Clustering Coefficient

The clustering coefficient was first proposed by Watts and Strogatz [94] as a measure of the cliquishness of a neighborhood. The clustering coefficient C_v takes values within [0, 1] and it actually expresses the probability of two adjacent nodes to a reference node v to also have a direct link. From another perspective, the clustering coefficient reflects how transitive the property encoded in the edges is for a node. In a friendship network, for example, it reflects the extent to which friends of an individual (node) v are also friends of each other. Consider that a node v has k_v adjacent nodes. If all its adjacent nodes were also connected with each other, they would form a clique of size k_v having $k_v(k_v-1)/2$ edges. The clustering coefficient C_v of v corresponds to the fraction of the allowable edges that actually exist and measures the local cohesiveness of v. The computation of the clustering coefficient actually involves the computation of the number of triangles in the graph, which has polynomial complexity. The clustering coefficient has been extended for weighted undirected graphs. The

following definition is introduced in [95] and can be applied in both binary and weighted representations of networks of weight values within (0, 1].

$$C_v = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{vi} w_{vj} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} w_{vi} w_{vj}}$$

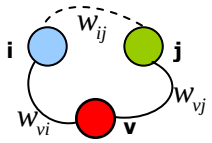


Figure 2.7 The clustering coefficient expresses the probability of two adjacent nodes of v to also connected with each other.

2.3.6 Assortativity Coefficient

Given a specific property to the nodes of the network, the assortativity coefficient expresses the degree of similarity or dissimilarity between adjacent nodes in the network reflecting the preferential attachment of the nodes of the network with respect to the specific property. Examples of node properties include age, race, a cultural trait and node centrality, which is the most commonly applied property. The assortativity coefficient is defined as the Pearson product-moment correlation coefficient between the properties of the nodes of every pair of the graph and takes values within [-1, 1], where -1 expresses perfect disassortative networks and 1 indicates perfect assortativity. The assortativity coefficient was initially proposed by Newman [96] and was applied in several network models and real networks in order to study their assortative mixing by degree. The work of Newman shows that biological networks in contrast to social networks tend to be disassortative indicating a preference of high-degree nodes to attach to low-degree nodes. The definition of the assortativity coefficient was extended by Leung et al [97] to characterize the degree correlation of weighted graphs. The weighted clustering coefficient thus measures the tendency of having a high-weighted link between nodes of similar degree.

For any scalar property $\rho : V \to \Re$ assigned to the nodes of the graph, the weighted clustering coefficient r^w is defined as follows [78]:

$$r^w = \frac{4H \sum_{\{u,v\}\in E} w(u,v)\rho(u)\rho(v) - \left[\sum_{\{u,v\}\in E} w(u,v)(\rho(u)+\rho(v))\right]^2}{2H \sum_{\{u,v\}\in E} w(u,v)(\rho(u)^2+\rho(v)^2) - \left[\sum_{\{u,v\}\in E} w(u,v)(\rho(u)+\rho(v))\right]^2}$$

where $w(u,v)$ corresponds to the weight value of the edge (u,v) and H is the total weight of all edges of the graph. If all the edge weights of the graph are equal then the weighted clustering coefficient reduces to the coefficient as defined for the binary graphs.

2.3.7 Network-level measures and statistical characterization

2.3.7.1 Network Centrality

Node centrality measures such as the degree or strength provide insight regarding the location and importance of a node in the graph. Network level centrality measures consider the centralities of all nodes of the graph. A network that is dominated by a few highly central nodes is highly centralized. If these nodes are removed the network rapidly becomes disconnected. The network centrality $C_x(G)$ is determined by the mean value of the

differences between the maximum centrality C_x^* over all nodes and the centrality of each node in the graph, where the index x represents whether the node centrality is measured with respect to strength or degree.

$C_x(G) = \frac{1}{n-2}\sum_{u \in V} C_x^*(v) - C_x(u)$, where v is the node that exhibits the highest centrality in the graph noted as C_x^*.

The network centrality takes values within [0, 1]. A star topology, in which at most one node has degree greater than one, has network centrality equal to 1. A complete graph, on the other hand, in which all nodes have the same degree, has network centrality equal to 0.

2.3.7.2 Degree and Strength distribution

The *degree distribution*, $P(k)$, expresses the probability that a selected node v has degree $d(v)$ equal to k. The probability $P(k)$ is estimated by counting the number of nodes $N(k)$ having degree k and dividing by the total number of nodes n.

The degree distribution is extended for the weighted graphs to the strength distribution $P(s)$, which corresponds to the probability that a selected node has strength s.

These distributions have been extensively studied in theoretical network models. The degree and strength distributions capture generic features of the binary and the weighted graphs respectively and have been used to distinguish different classes of network topologies. The distributions can provide information regarding the existence of a characteristic degree or strength in the network or the identification of highly connected nodes, known as hubs. The degree and strength distributions have proved of great importance in studying real networks, such as social networks, the internet and biological networks.

2.3.7.3 Clustering spectrum

The average clustering coefficient $C(G) = \sum_{v \in V} C_v / n$ measures the global density of interconnected triplets in the graph and expresses the overall tendency of nodes to participate in clusters.

The clustering coefficient can be also studied as a function of the degree k and the strength s for binary and weighted graph representations respectively. The degree-dependent average clustering coefficient $\overline{C}(k)$ is defined as the average clustering coefficient for nodes with degree k [74, 97].

The function $\overline{C}(k)$ quantifies the node's centrality with the degree of modularity as measured by the clustering coefficient and has been used in order to provide insights regarding the organizational structure of the network. In networks of hierarchical architecture, it was shown that clustering coefficient of a node of degree k follows the scaling law $\overline{C}(k) \sim k^{-1}$ [76]. The power law scaling of the clustering coefficient with the node degree can be considered as a signature of hierarchical organization and modularity in networks and has been observed in the metabolic networks of many organisms [98, 99].

2.4 Theoretical Network models

Network models have been developed in an attempt to produce graphs whose properties reproduce those of real data, analyze the mathematical properties of networks theoretically and also explain the evolution processes that generate the characteristics in structure of the real-world networks. Not all real networks exhibit similar properties and statistics. However, it is surprising that in this diversity of networks the underlying principles of their architecture is common to most networks. Knowledge of the properties of network models allows a better understanding of the complex organization, the local and global properties of the networks in nature. Furthermore, understanding the evolutionary origin of the observed networks is important for a better understand of the systems. It is an open question what type of growth rules including structural and functional mechanisms could explain the observed networks. The models presented here have a direct impact on the understanding of biological networks. The models concern binary networks with undirected edges. A nice review about the evolution of networks from the statistical physics point of view can be found in [76]. This section is provided for completeness.

2.4.1 Erdős–Rényi (ER) Random graph

Different random processes produce different probability distributions on graphs. The simplest and most well-studied random network model with undirected edges was introduced by Paul Erdős and Alfréd Rényi (ER model) [100]. The ER random graph is obtained by starting with a set of n nodes (fixed) and connecting each pair of nodes with probability p, which is random and uniform.

On average, this network has $pn(n-1)/2$ edges. The node degrees follow a binomial distribution, which indicates that the average degree $\bar{k} = (n-1)p$ and that most nodes have the same number of connections (the average degree). For large n the distribution approximates the Poisson distribution. The average shortest path $\bar{l}$ of the networks is estimated to be $\bar{l} \sim \log n / \log(np)$. Thus, the random network is characterized by the small-world property. The clustering coefficient is independent on the node degree and the assortativity coefficient approximates 0 for large graph sizes.

The weighted counterpart of the Erdős- Rényi random graph model (WRG) was proposed by Garlaschelli [101]. Many of the mathematical properties of the model were derived exactly. The model is characterized by a geometric weight distribution, a binomial degree distribution and a negative binomial strength distribution.

2.4.2 Scale-free graphs

2.4.2.1 Barabási–Albert (BA) model

The degree distribution $P(k)$ of many real-world networks is observed to follow scale-free distribution; that is $P(k) \sim k^{-\gamma}$, where the degree exponent γ varies between 2 and 3. The term scale-free is rooted in statistical physics and indicates the absence of a typical node of characteristic scale in the network contrary to random graphs in which the nodes have degree around the mean degree of the network. The power-law scale that characterizes this

distribution indicates that highly connected nodes have a statistically more significant chance of occurring than in the ER random graph.

The Barabási–Albert model [102] was proposed in order to explain the way networks self-organize into scale-free structures while growing. The model introduces the concepts of growth at which the number of nodes increases over time and of preferential attachment, which is expressed as a linking preference to nodes of high degree. Growth and preferential attachment are two generic features that are observed in most real networks and which were not incorporated in the random model described previously. Preferential attachment can be seen as an example of the 'rich gets richer' phenomenon, a positive feedback where random variations are magnified throughout evolution process.

In the BA model, the growth begins with an initial configuration consisting of a small number of nodes n_0, which must be at least two of degree at least 1. If these initial conditions do not stand then the network remains disconnected throughout the whole evolutionary process. At every time step, a new node with $m \le n_0$ links is added. The preference function is expressed as follows. The new node is connected to an already existing node i with probability Π_i which is proportional to the degree k_i of the node i; that is $\Pi_i = k_i / \sum_j k_j$, where the sum is over all already existing nodes. After t time steps the network consists of $n_0 + t$ nodes and mt edges. The degree distribution of the evolved network exhibits power-law scaling with degree exponent $\gamma_{\text{model}} = 2.9 \pm 0.1$ independent of time t. It is shown that the degree of a node increases with the square root of time. Nodes that appeared early on the evolution history of the network exhibit the highest degrees. The mean path length $\bar{l}$ of the scale-free networks with exponent $2 < \gamma < 3$ follows $\bar{l} \sim \log\log n$. These networks thus exhibit the ultra-small-world property. The BA model is not characterized by any inherent modularity and therefore the clustering coefficient is independent of the node degree. As the number of nodes increases it was shown that the BA model becomes assortative neutral meaning the assortativity coefficient approximates 0.

Although growth and preferential attachment are reasonable assumptions for many real networks there are situations where these concepts do not hold. Different generative models capable of also producing scale-free degree distributions such as the copying and the fitness models have been introduced [103-106]. Furthermore, it was shown that a simple process on weighted ER graphs is also capable of generating scale-free networks [107].

2.4.2.2 Copying model

The growth process of the copying model [105] was inspired by the World Wide Web and the way a new web-page is usually added to the system. The network grows by replicating nodes and their corresponding edges allowing some tolerance on the replication procedure. Specifically, at every time step a node chosen at random is duplicated. From the m_0 connections of the original node, the duplicate node keeps each connection with probability $1-a$ or it is rewired with probability a. Since, at each time a node is selected at random, the probability of a duplicate node to be connected with a node is proportional to each degree. In other words, highly connected nodes are more likely to gain new connections through the evolution process. This growth mechanism can thus be considered a subtle version of the preferential attachment rule and emerges from the local rather than the global knowledge of the system. The duplication mechanism plays an important role in genome evolution and might describe the evolutionary origin of the scale-free topology observed in cellular networks [74, 103].

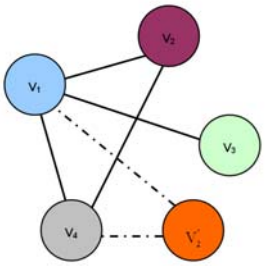


Figure 2.8 Copying model. A node v_2 is selected at random and copied. The replica of v_2, the node $v_2^{'}$ is adjacent to the same nodes (v_1 and v_4) as the node v_2.

2.4.2.3 Fitness model

The fitness model [104] is based on the assumption that two nodes are connected with each other when there is a mutual benefit, which depends on intrinsic properties of the nodes. The intuition behind this assumption is that in many real cases fitness comprises an inherent competitive factor capable of determining the network interactions. Fitter nodes gather more connections contrary to less fit nodes.

The network model is reconstructed as follows. Each node i of the network is assigned with a real number x_i measuring its importance ("fitness"). The value of the fitness is a random number taken from a given probability distribution $\rho(x)$. An edge is assigned between a pair of nodes according to the *linking function* $f(x_i, x_j)$, which depends on the fitness values of the corresponding nodes. The reconstruction procedure of the model as described is static. However, new nodes can be added and attached with the other nodes in the network according to the fitness rules. If the linking function is constant, uniform and equal to p then the model reduces to the ER random model. The linking function is of the form of a Heaviside step function representing processes where an edge is assigned if a function of the fitness values (such as the sum) is above a given threshold. It is shown that several fitness distributions are capable of generating scale-free degree distributions [103, 108].

2.4.3 Modules and Hierarchical graphs

Beside the scale-free degree distribution that characterizes most real systems, a high mean clustering coefficient is also evident in many real networks from social to biological networks. It is also observed that the clustering coefficient is at a high degree independent of the network size. The clustering coefficient is a measure of the degree of modularity that is present in a given network. Modules are subgraphs which by definition are characterized by a dense internal connectivity whereas they are relatively isolated from the rest graph making their existence unlike in scale-free topologies. Thus, the question that arises is the following: how a scale-free network can exhibit high degree of modularity. Ravasz and Barabasi [109] suggested that in order for clustering and hubs to coexist, modules should connect with each other in a hierarchical manner generating the so called *hierarchical network*.

The deterministic hierarchical network model is reconstructed as follows (Fig. 2.9). The initial configuration consists of a clique of size 5 representing the characteristic module of the graph. All the nodes of the clique are equally important. Among them one is selected to play the role of an internal node of the module whereas the rest become peripheral nodes. Four replicas are then generated. The peripheral nodes of the replicas are connected with the internal node of the initial module, which is now becoming a hub. The graph now consists of

25 nodes. Four replicas of the 25-node module are generated in the next iteration. The 16 peripheral nodes of its new replica are connected again to the internal node of the initial module generating a 125-node module graph. In this manner the replication and connection steps can be repeated indefinitely.

The hierarchical network reconstructed in the way previously described displays scale-free degree distribution of degree exponent $\gamma = 1 + \ln 5 / \ln 4 = 2.161$. Numerical simulations show that the mean clustering coefficient is independent of the size of the network and is approximately equal to $\overline{C} = 0.743$. The hierarchical structure of the network is quantitatively captured in the degree-dependent clustering coefficient $\overline{C}(k)$ which is observed to follow the scaling law $\overline{C}(k) \sim k^{-1}$. This observation implies that hubs are not part of highly connected clustered areas however they play the important role of connecting the different clustered areas into a single, integrated network. A stochastic reconstruction model has also been proposed and analyzed in the work of Ravasz and Barabasi [109].

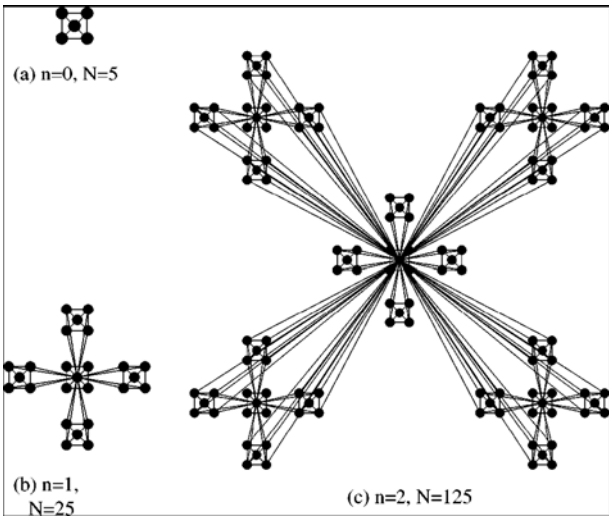


Figure 2.9 A deterministic hierarchical network reconstruction. The illustration is borrowed from [109]. The reconstruction begins from a clique-module of size 5 (a). In the next step (b) for replicas are created and the peripheral nodes of each are connected with the internal node of the original module. Four replicas of the 25-module graph are created and connected to the original (c). The diagonal edges within the original module and its replicas exist but are not shown.

"What I cannot create I do not understand"
Richard Feynman

3. Modeling biological systems

3.1 Introduction

Biological systems such as cellular systems generally consist of a large set of diverse and multi-functional components that interact dynamically and selectively in response to certain environmental conditions and signals. The biological systems usually obey in the conservation laws of mass and energy whereas the interactions between the components are governed by a set of biophysical laws and constraints, which have been shaped under a variety of evolutionary, ecological and environmental forces.

The identification of the individual components responsible for a particular biological phenomenon (**Reductionist approach**), is important but unable to explain the emergent properties of the whole system. Understanding how the various biological components that constitute a biological system interact as a whole in time and space is crucial for deciphering its functions and behaviors. Towards this goal, **Systems Biology** approaches are frequently used to identify the components of the whole system and their relationships as well as to build *in silico* models capable of integrating the available biological information and data, representing graphically the inherent complexity so that it becomes more comprehensible, explaining the complex biological processes and providing testable predictions.

Given any biological information, which is known *a priori* about the system and the available data the system produces, inference of the interaction map known as network or structure of the system (structural inference) is the first task modeling is usually concerned about. Once the network of the biological system under study is reconstructed, methods capable to explore its structural properties and organization can be applied. Methods of the network-based analysis of the system were analyzed in the previous chapter. Modeling also concerns inference of the mathematical formalisms including the involved parameters, which approximate the dynamic laws and constraints the system follows (dynamical inference).

Identifying the mathematical framework that better describes a biological system however is not trivial. The information, which is available for the system under study as well as the type of information that is important to infer from the model, define more or less the golden section between a coarser to a finer level of abstraction and determine the selection of the appropriate method. The current biological knowledge of the system under study certainly

places limits to the level of description and the challenge is thus to build mathematical models with the available data at hand that will be able to reveal new properties and understand the system functioning to further guide new experiments. Modeling is subject to an iterative refinement procedure which terminates when a valid and biologically plausible, unified description of the system is found.

3.1.1 Structural inference

The behavior of a system can be decomposed into the orchestrated activity of the components that interact with each other through pair-wise interactions. The selective interactions among the components of the system define a connectivity map. These interactions reflect a sort of influence that (the value of) a component has on the (values of) other components without any definition of any timings of these interactions.

Based on various assumptions, the network-based inference algorithms attempt qualitative describe the topology of a biological system (structural inference). Among all possible structures, a network which best describes the information given for the underlying system is searched. The inference methods can be distinguished by the way they remove influences of other components from the observed correlations. The inference algorithms range from discrete deterministic models such as Boolean networks to probabilistic models such as Bayesian networks.

The network (or structural) inference problem can be formulated as follows:
"Given a set of interacting components as well as any prior knowledge and any set of observations that the system components produce, find the network connectivity that satisfies a given set of constraints and assumptions."

3.1.2 Dynamical inference

The dynamical inference problem aims to identify how the components of the system vary over time in response to external, internal and physiological cues and determine its functional states. Therefore, it mainly concerns the determination of the biophysical description of the system, including assumptions and decisions such as the deterministic or stochastic nature of the system dynamics, the influence of the homogeneity of the actual reacting volume in the dynamics, and the linear or non-linear dynamic interrelation of the system components. The optimal model is selected based on the available data, and the biological knowledge and assumptions that are most proper to fit to the biophysical dynamics of the system under study. In an attempt to make the biophysical description of the system specific the dynamical inference problem is also concerned of identifying the missing parameters involved as well as incorporating them to the pre-defined mathematical framework. Actually many of the kinetic parameters (reaction rates, diffusion or transport rates) involved in the governing equations are usually missing. Therefore, methods for estimating those parameters from experimental data should be applied. Optimization of the parameters involved is a problem of its own and usually a large computational effort is needed.

The most common description of the dynamical inference problem is the following:
"Given sequential snapshots of the time evolution of the components of a system (time-course data), a function that explains the dynamic change of the values of each component with respect to the system's past values is asked. The solution for the problem should explain and predict the population levels of the species at any time, given the molecular populations of the biochemical reacting system at an initial time."

3.2 Intracellular biochemical systems

The central dogma of biology states a rather simple principle: DNA, the carrier of the genetic information is first replicated and then it is transcribed to messenger RNA (mRNA), which is thereafter translated into proteins. Once information gets into protein, it can't flow back to nucleic acid. Behind this simple idea of genetic information flow, lays a dramatically complex cascade of regulatory and chemical events.

The complexity of cell systems stems from the diversity, multi-functionality and plethora that characterize their components. The set of the intracellular components (genes, gene products, metabolites) interact selectively and usually non-linearly in response to internal and environmental signals to produce coherent rather than complex behaviors and sustain the characteristic features of life such as growth, cell division, intracellular communication, movement and responsiveness to a variety of cues. Cell components can also be seen as players in a 'survival' game with short and long-term goals that follow certain strategies with the potentiality of updating and self-adjusting. These strategies involve the temporal succession of the biochemical interactions that the components of the system will undergo in order for the system to accomplish certain functional tasks towards maintenance of life. The biophysical laws in which the system obeys as well as the evolutionary process that has driven the system towards a better survival and source utilization are actually the forces that shape the regulatory strategies. Thus, the system efficiently orchestrates its components and realizes its tasks.

The way of shedding light into the complex cellular system to further build *in-silico* models capable to explain complex cellular processes, make testable predictions and resemble real life includes the identification of the biological components that are involved in cellular functions, the determination of their interactions as well as the obtainment of expressions that quantitatively and precisely describe every little detail and principle of the system (figure 3.1). Towards this goal, looking the system globally rather than locally is very important.

To reduce the complexity the cellular system is usually decomposed into interconnected functional sub-networks (known as functional modules [110]) whose task is separable from those of other modules. However, even functional modules are context dependent and open subsystems and as such should be taken into account in modeling and predictions. Furthermore, we should always bear in mind that the data produced by the system are the result of global scale interactions that take place in many levels from genome to phenotype.

3.2.1 Current data

Interactions between molecular components in a cell are fundamentally described by chemically reactions. The cell can be seen as a chemically reacting system in which functions rely on the interactions among its chemical constituents.

Researchers in molecular biology attempt to understand the interactions between DNA, RNA, proteins and other chemical components of cells. They focus on the generation of information regarding individual cellular components, their chemical composition and the biological functions in which they are related to.

Over the last decade, technological advances allow the detection of thousands of biological components simultaneously providing information about the chemical composition of cells at genome-scale and forcing biologists to view cells as systems. These high throughput techniques also known as *omics* data sets [111] include the entire DNA sequencing and annotation which is available for a growing number of organisms (genomics), the DNA microarrays which can be used to measure changes in gene expression levels of the whole genome in parallel (trascriptomics), protein abundance measurements, their interactions and functional states (proteomics), measurements of the metabolite profiles, their presence and

concentration based on various spectroscopic techniques (metabolomics) and measurements of the metabolic fluxes (fluxomics). Flow quantities are related to transport or conversion phenomena. Flux represents the passage of molecules of a particular metabolite through a metabolic or transport step per unit cell mass per unit of time. Efforts have been also made towards the identification of the physical, sub-cellular location of all proteins in the cell (localizomics). The omics data describe the biochemically reacting network of the cell at a given time or condition with respect to its components, interactions and functional states (Figure 3.1).

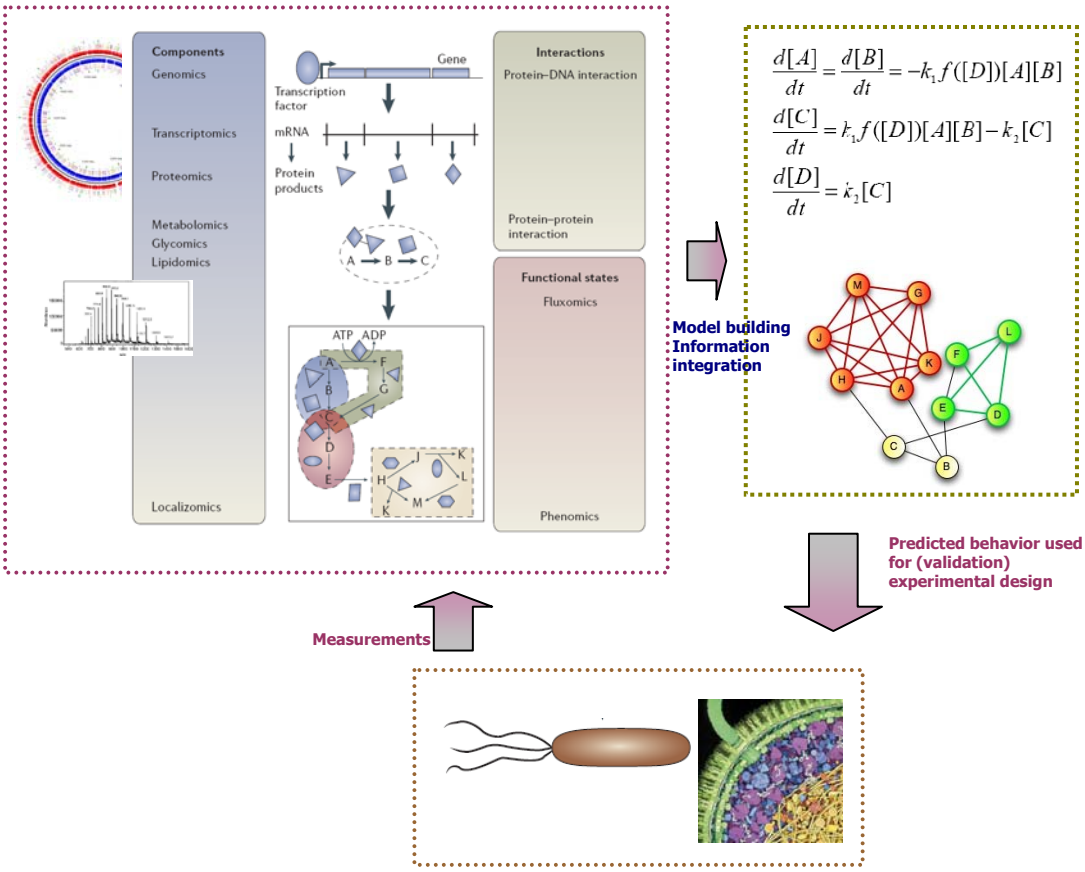


Figure 3.1 The iterative modeling procedure integrates all the available 'omics' datasets for a given organism and determines the interactions between its components as well as their dynamics in an attempt to reveal the overall cellular phenotype. Adapted from [111].

3.2.2 Challenges

The challenge in the post-genomic era is to properly integrate all the available information so as to reconstruct, as accurately as possible, the complex network of interactions inside a cell, beginning from the identification of its static structure and moving on to characterize its dynamic behavior and functional role. In order to cope with the massive amount of data that need to be analyzed and synthesized for modeling a biological network, mathematical and computational methods are inevitably required. The datasets are usually noisy and highly heterogeneous and in some cases incomplete with respect to their insufficiency to detect all the components involved or cope with the system dynamics due to low sampling rate. The quality of data affects the design of models that aim to reconstruct cellular systems. Novel computational methods developed specifically for biological data are essential. The main properties of the system, such as complexity, robustness and adaptation must first be derived from available data, in order to design a model that not only describes faithfully its structure but also makes useful predictions about its behavior.

The cellular interacting system is usually decomposed into signaling, regulatory and metabolic network [17]. Signaling networks are responsible to transmit signals from the cellular environment to the nucleus or other cellular compartments. Transcription regulatory networks rely mainly on protein-protein interactions and DNA-protein interactions determining which genes are expressed under given environmental and signaling cues. mRNAs, transcription factors, microRNAs and epigenetic modifications are combined to form a complex regulatory network allowing cell to respond to a changing environment. Transcriptional level regulation is of great importance particularly in higher organisms. Metabolic networks comprise the set of the enzymatic reactions responsible to transform substrate molecules into essential building blocks and other vital molecules essential for cellular growth and maintenance. Metabolism comprises one of the most fundamental and well-conserved cell processes of living organisms. However, most of these chemical reactions are thermodynamically unfavorable and for that reason they require the presence of the appropriate enzymes to take place. Regulatory mechanisms interacting with the cell's internal and external environment are responsible to provide to the cell the proper enzymes at the proper times making metabolism subject to regulation.

These networks are not independent and an integration of all the networks into a complete genome-scale reconstruction of cellular functions remains to be established. Genome-scale metabolic networks at the level of individual chemical reactions have been reconstructed for microorganisms. On the other hand, signaling and transcriptional regulatory networks are mostly described by causal relationships rather than chemical equations. Genome-scale description of these subsystems is not provided since the biophysical reaction mechanisms are not yet fully understood.

Quantitative information on kinetic parameters and molecular populations is also limited, making the model reconstruction harder. In metabolic networks for organisms like the bacterium *Escherichia coli* or the yeast *Saccharomyces cerevisiae*, more knowledge is available but the data is far from being fully complete. An exceptional case comprises the human red cell for which all the kinetic parameters are available [112, 113].

3.2.3 State of the Art

Boolean [114, 115] and Bayesian [116-118] networks comprise the main methods of the network-based algorithms, that qualitatively describe the topology of a biological system (structural inference). Based on various assumptions, these algorithms search for a network, among all possible structures, which best describes the information given for the underlying system.

Boolean networks have frequently been used to infer and model the transcriptional regulatory network from sets of time-course gene expression data. A variety of algorithmic strategies for inference have been proposed in the literature [119-121] and applied for the Boolean network reconstruction and analysis of the Saccharomyces cerevisiae cell cycle [122, 123], the immunology microarray data-sets [124] and the metastatic melanoma related gene expression data [125, 126], among others.

Bayesian networks have several advantages such as they can handle hidden variables, missing data and unobserved values of system components by allowing latent components to be present in the network; they are robust for noisy data, as they are probabilistic in nature, and are inherently optimal for describing processes composed of locally interacting components. Bayesian methodology provides a principled way of incorporating additional information as prior knowledge, but the challenge to assign a weight of trust in these different sources of information remains [127-130]. As dynamic data are richer in information than static data, their use should further reduce the ambiguities concerning the underlying network structure. Temporal or dynamic Bayesian networks are extensions of the Bayesian

networks that explicitly model the stochastic evolution of the system components through discrete time.

Under the deterministic framework, the biophysical laws that describe the temporal evolution of a system produce precisely and consistently the future states of the system given its initial condition. A variety of mathematical expressions varying from linear [131-136] to piece-wise linear [137-139] and nonlinear models [140-143] have been proposed in the literature to specify those laws and explain the observations.

On other hand, stochastic phenomena have been also observed in biological systems. Even in cloned cell populations and under the same (as possible) experimental conditions, significant phenotypic variations have been reported in each cell including variations in the rates of development, morphology and population levels of each species [144-146]. In systems where the populations of the species are large enough (e.g., metabolites), random effects are averaged out. Nevertheless, if the population of a system component is low enough (e.g., mRNA copies, transcription factors), the fluctuations in the molecular levels (e.g., protein levels) could not be predicted by a deterministic approach. When the microscopic fluctuations produce macroscopic effects [32] that the deterministic reaction rate equations are unable to predict, microscopic stochastic simulation approaches described in a later section are required [147].

All organisms, even bacteria, show a spatial organization into cellular compartments. In the modeling however a homogeneous distribution of the cellular components is commonly assumed. Including a spatial constituent into the biochemical dynamic expressions is essential when different reactions evolve differently in separate compartments and the molecular mobility causes significant non-linear delays to the dynamic system. Spatial phenomena arise when reaction rates are shown to be comparatively faster than diffusion rates. Spatial phenomena are also apparent when the reacting volume is of high molecular density (molecular crowding). Signaling pathways have reported suffering from phenomena such this [148-150].

In principle, with enough computing power, an initial picture of the system at the molecular level and a proper unified theory of the biophysical laws that govern the system's dynamics, which takes into account all peculiarities of the biological system (in homogeneity, randomness, redundancy, robustness), a whole cell simulation at that molecular-reaction level of detail could be achieved. However, even though a massive amount of experimental data is currently available and substantial biological knowledge has been gained, they remain insufficient for the inference of the missing knowledge, in order to simulate large scale systems at molecular resolution. There are compromises that, if properly applied, may improve the simulation speed and reduce the dimensionality problem and the parameter space, while making minor sacrifices in the description accuracy of the phenomenon. For example, models that partition the system into subsystems, where different assumptions can be applied, have been proposed including the stochastic-deterministic hybrid models [151, 152]. A recent multi-level integrated software tool for simulation of complex biochemical systems is COPASI [153]. COPASI incorporates deterministic and stochastic approaches to simulate biochemical reactions and also provides a package of optimization algorithms to estimate the unknown parameters involved. To keep the problem tractable, the cellular system has also been partitioned into functional modules where detailed kinetic models can be constructed for each. Those divide-conquer (bottom-up) approaches face two difficulties. The partitioning of the system into functional or mathematical parts, in addition to the integration which has to be followed is not always a trivial task. Furthermore, when validation or optimization is needed for the sub-models, we must have in mind that the data are usually referred to the complete system and not to the parts which are indeed not independent of the rest system. Based on the modular approach, nice examples of integrative dynamic models are presented by Snoep et al. [154] and Klipp et al. [155]. Alternative models, which simulate large scale systems as a whole by incorporating information and data from genes to proteins and enzymes, have also been proposed, sacrificing dynamic description resolution.

Constraint-based models [17] are widely used as top-down models, for the investigation of the metabolic capabilities of certain organisms under specific environmental conditions and perturbations. Dynamic phenomena can be approximated by changing the constraints to shape according to the feasible flux space. The temporal path of flux distributions that the system undergoes throughout its dynamic evolution remains an open problem. Additionally, a better way to incorporate other interacting systems such as signal pathways, and gene regulatory networks to the complex metabolic network leaves room for improvement towards a multi-level integrated system.

In the following focus will be given in the metabolic network reconstruction and the constraint-based framework for its functional analysis. A brief description of bacterial growth and its simulation based on the constraint-based analysis is presented.

3.3 Genome-scale Metabolic Network Reconstruction

Cells of all living organisms need to find in their environments the substances required for energy generation and biosynthesis. The energy sources, that are available in the environment, are converted into ATP and other cellular building blocks to construct essential components such as proteins and nucleic acids in order for the cell to maintain growth and survival. This process is called metabolism and comprises one of the most fundamental and well-conserved cell processes of living systems. Available high-throughput data allow the description of cellular metabolism at a high resolution, which considers the set of all biochemical transformations that take place within the cell.

The reconstruction of a genome-scale metabolic network [43] for an organism under study relies on assembling various sources of information such as biochemical data, genome-sequencing, physiological data as well as simulation data. The confidence levels of the information sources highly differ; with the biochemical data to provide the highest accuracy and strongest evidence for the presence of a chemical reaction in the network. However, the information sources can be carefully integrated in an iterative manner which gradually develops the description of the biochemical system. The reconstruction process ultimately results in the generation of a biochemically, genetically and genomically structured (BiGG) database [156], which can be further utilized both computationally to predict phenotypic properties and biologically to guide new experiments with the aim to bridge the genotype-phenotype gap and provide an understanding of the functions of a living cell. For example, the advantage of looking at the whole picture of interactions in a cell is evident when the function of a gene product is unknown but its presence can be inferred based on the inability of the cell to function without it. Thus, during the network reconstruction process formerly un-annotated gene functions are incorporated into gene-annotation knowledge by analysis of incomplete but essential metabolic pathways. This process is known as gap analysis and comprises a significant contribution of the reconstructed metabolic network to biological knowledge.

The metabolic network reconstruction process has to give answers to several questions including whether an enzyme is present in the specific organism, which are the reactions it catalyzes, if there are cofactors involved, what is the stoichiometry of the reactions, whether a reaction is reversible or irreversible and where are reactions localized within the cell. For modeling purposes, the biomass composition of the organism, which includes the set of all precursors and building blocks that are essential for biomass production, has to be defined as well. In general, the organism-specific biomass composition can change under different growth conditions, but most of the times it is considered constant as it merely changes the simulation results [157].

The iterative, ongoing reconstruction process is usually non-automated and includes four main steps. The initial reconstruction is built manually upon gene-annotation data coupled with information, which links known genes to functional categories. Genome annotations are

subject to revision and updates. Online databases such as KEGG [158, 159], ExPASy [160], BioCyc [161] are commonly used at this stage. For the reconstruction of a high quality and well-curated network, a thorough examination of the organism-specific primary literature and available biochemistry textbooks are very important. As a next step, the reconstructed network is converted into a mathematical model and analyzed under the constraint-based framework, which is analytically described in the next section. The predictions of the model are compared with currently available experimental data for the organism under study, which can include physiological data such as metabolic by-product secretion and growth, gene essentiality data and growth perturbation experiments. Thus, whether the reconstructed network is capable of reproducing observed cellular behaviors is determined (validation step). Under that respect, the reconstructed network is subject to continued wet- and dry-lab cycles, which improve its accuracy and permit the investigation of new hypotheses.

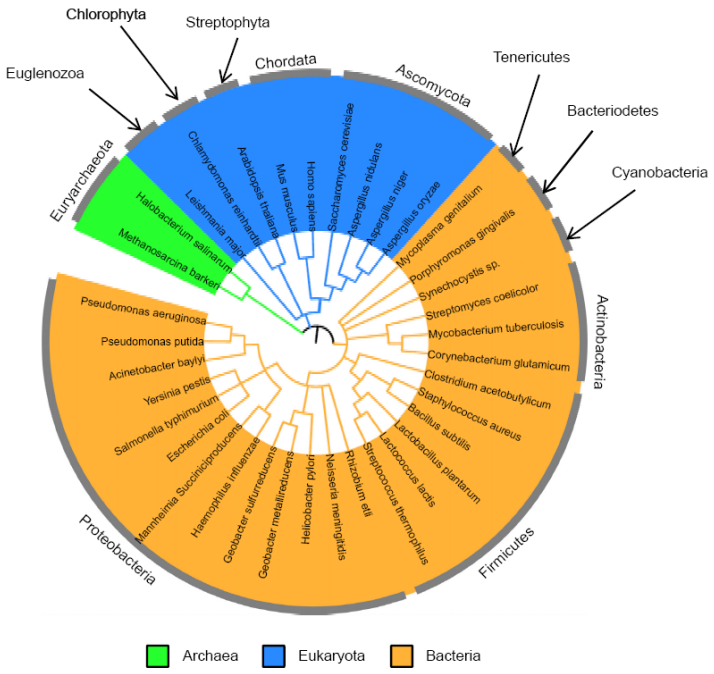


Figure 3.2 The phylogenetic tree of all the species for which the genome-scale metabolic networks have been reconstructed. The figure is borrowed from [43].

The first genome-scale metabolic model was reconstructed in 1999 for the organism *Haemophilus influenza* [162], a Gram-negative, rod-shaped bacterium responsible for many human diseases. Today, more that 50 genome-scale reconstructions (Fig. 3.2) have been published [43]. Among them, bacteria [44] comprise the most represented domain with the bacterium *Escherichia coli* gaining the most attention as a model organism [53, 163]. In fact, the history of reconstruction of *Escherichia coli* begins in 1990 [164]. On the other hand, baker's yeast, *Saccharomyces cerevisiae*, comprises a eukaryotic model organism with high industrial importance, which has been intensively studied as well [50, 165]. Genome-scale metabolic reconstructions for plans are missing with only exception *A. Thaliana* [166].

As it is already pointed out metabolic networks do not operate independently; they communicate with other cellular processes such as the transcriptional regulatory and the signaling networks (Fig. 3.3). Complex regulatory networks must be in place to ensure the proper coordination of all the biochemical events that allow cell functioning. Even in bacteria optimal growth requires coordination of the cell cycle with metabolic processes. In addition, cellular phenomena including protein modification, motility, adhesion as well as fate processes like mitosis and apoptosis, which occur in multi-cellular organisms are simultaneously observed within a cell converting the system to highly complex, stochastic and spatially dependent. Small-scale integrations of the cellular sub-networks have been developed [167, 168]. A Boolean representation of the transcriptional regulatory network

(TRN) has been incorporated into the genome-scale metabolic model and is presented in the following section. A merging of the genome-scale metabolic, regulatory and signaling networks represented at high resolution into a common framework remains however a challenge [43, 44]. Nevertheless, the metabolic network reconstruction comprises a natural, first, modeling step towards understanding the synthetic capacity of a cell and its metabolic capabilities.

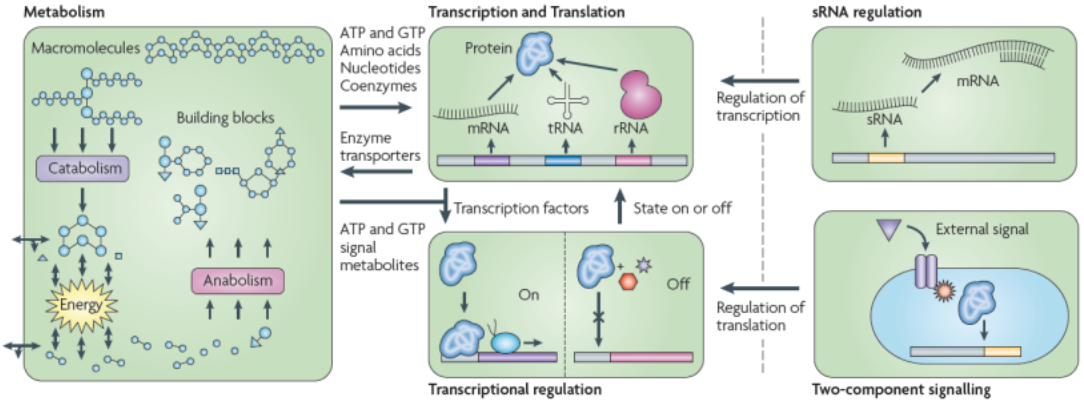


Figure 3.3 Network integration [44] of metabolism, transcriptional regulation and translation (TRN), two-component signaling pathways and translational regulatory network controlled by small RNAs. Both signaling pathways and sRNA regulatory network are not completely known even in *E. coli* and their integration in the whole-cell reconstruction comprises an open challenge. Boolean representations of TRNs have been incorporated. The figure is borrowed from [44].

3.4 Genome-scale Metabolic Model

3.4.1 Stoichiometric Matrix

Chemical reactions link the molecular components found in the cell and form a network. The question is how from the complex network of the biochemical interactions we can determine cellular phenotypic functions.

The metabolic network can be represented by a stoichiometric matrix S (Fig. 3.4), which includes the stoichiometric coefficients of all the reactions that comprise the network [17]. Every column of S corresponds to a reaction and every row corresponds to a metabolite. The number of the columns of S is greater that the number of rows because the set of the biochemical reactions that the metabolic networks consist of is always larger than the set of metabolites. The entries in the matrix are the stoichiometric coefficients of the corresponding reactions, which are integers expressing the relative proportions of the compounds involved in a chemical reaction. Thus, the matrix S represents biochemistry (elemental balancing) but also provides information about how the reactions are interconnected.

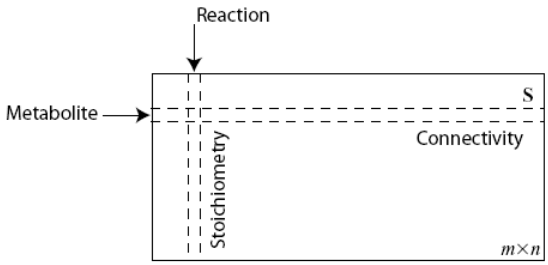


Figure 3.4 A graphical representation of the stoichiometric matrix [17].

Metabolic networks can be represented as flow networks. In fact, the metabolic network represents the channels for the flow of material and generation of Gibbs free energy, which are constrained by the stoichiometry as well as the conservation laws of mass and energy. The conservation of mass expressed as mass balance is defined in terms of the flux through each reaction channel and the stoichiometry of that reaction. A set of coupled differential equations can be formed to express the temporal evolution of each metabolite in the system with respect to the conservations laws. In that way, the set of the chemical reactions that comprise a network can be represented as a set of chemical equations.

The stoichiometric matrix S transforms the flux vector $\vec{v}$ into a vector that corresponds to the time derivatives of the concentrations of the metabolites $\vec{x}$ involved in the network. This equation (Eq. 3.1) represents the fundamental equation of *dynamic mass balances.*

$$\frac{d\vec{x}}{dt} = S\vec{v} \qquad \text{Equation 3.1}$$

Depending on how the system boundaries are drawn, the reactions and correspondingly the fluxes through the reactions can be partitioned into internal and external. It is common to draw the systems boundary around the cell because it is consistent with physical realties. Thus, beside the internal set of fluxes, fluxes entering and leaving the boundary of the system (exchange fluxes) through the transport mechanisms are present as well. However, it might be the case that different subsystems of the network are governed by different dynamics and in that case virtual boundaries have to properly be chosen.

3.4.2 Constraint-based framework

Equation based models attempt to reveal the real biophysical forces and describe the dynamic laws that drive a cellular system to its functional properties by going deep into even microscopic scales. In a complex system of such a variety of components and interactions, robust and highly regulated, as the biological systems are, the more plausible way to understand the system at that high resolution was to restrict studies to small parts of the puzzle and then follow the bottom-up approach towards the whole picture reconstruction. Detailed kinetic models are available for specific pathways and metabolic sub-systems [169-171]. However, genome-scale kinetic information regarding reaction rate constants, enzyme concentrations and metabolite concentration is missing. Current knowledge certainly places limits to the level of description and the challenge is to build mathematical models able to reveal new properties of the system and better understand its functions with the available data at hand in order to guide new experiments for further verification and extension of the models.

Constraint-based models consider the cell as a complete system that orchestrates its components under physiochemical constraints towards the accomplishment of metabolic requirements. The constraints that the system inevitably obeys include mass balance, energy balance and flux limitations of which not all are fully known. In fact, the constraints describe the biophysical laws that the components of the system obey, as also the evolution process that has driven the system towards a better survival and source utilization.

Constraint-based metabolic models utilize the genome-scale metabolic network reconstructions with the aim to integrate knowledge at different levels in the cascade from genes to proteins and further to chemical reactions and metabolic fluxes to describe and understand the overall cellular functions [47, 172]. The constraint-based framework allows the direct correlation between the genomic information and metabolic activity at flux level and the elucidation of properties (such as network robustness, product yield, and metabolic

versatility, environmental and genetic phenotypic effects) that cannot be described by descriptions of individual components.

The core assumption of constraint-based models is that the system reaches a *steady state* (intracellular flux balancing) that satisfies the physiochemical constraints under any given environmental condition (Eq. 3.2). The hypothesis is based on the fact that the time-constants which describe metabolic transients are fast (on the order of milliseconds to tens of seconds) as compared with the time constants associated with transcriptional regulation (generally on the order of few minutes or slower) or cell growth (on the order of hours to days). Under that macroscopic time scale, the cell can be considered as being in a quasi-steady state.

$$S\vec{v} = \vec{0} \qquad \text{Equation 3.2}$$

Originally, the models assumed that the metabolic system reaches a steady state constrained by its stoichiometry (mass balance). Nevertheless, the stoichiometry alone only bounds the solution space determining the region of the possible fluxes the system can reach (*feasible space*). The metabolic networks always have more fluxes than metabolites. Thus, the system is undetermined under the stoichiometry constraint alone. Thermodynamic constraints that determine the reversibility of the reactions involved and enzymatic capacity constraints were also included to place limits on the range of possible fluxes. Figure 3.5 abstractly depicts the successive restrictions of the flux space the constraints of the system impose.

The feasible flux space can provide valuable information about the functional states of the metabolic network. Convex analysis has been used to enumerate the unique set of all distinct metabolic routes (elementary modes) of the network [173-175] as also to determine the minimal set of convex vectors (extreme pathways) needed to describe all allowable steady state flux distributions [176-178]. Elementary mode analysis and extreme pathways comprise an important step towards the characterization and understanding of the solution space. What remains unanswered is the solution the cell chooses under the given conditions (observed phenotypes).

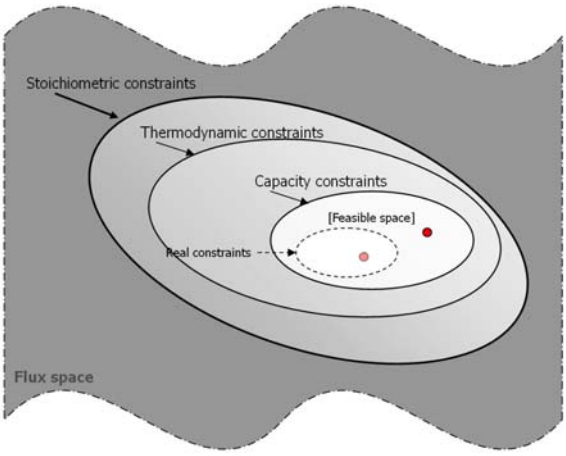


Figure 3.5 Constraint-based modeling. Biologically meaningful constraints applied on the metabolic system reduce the space of possible flux distributions. The observed flux distribution of the metabolic system operating at steady state should lie in the white area that represents the only feasible space under the given constraints. However, the real constraints may differ from those imposed by the model resulting to different solutions as depicted with the red and pink circle.

3.4.3 Primal Flux Balance Analysis

The Darwinian theory of evolution considers that the process of successive mutations and selections eventually lead an organism to a permanent state of phenotypic properties which are optimal to given environmental conditions. Thus, biological systems can be considered as the outcome of an optimization process.

Constraint-based models further assume that under given conditions the performance of a cell follows an optimization strategy in order to accomplish cellular tasks. The most commonly applied optimization functions include the maximization of biomass production, the minimization of nutrient utilization, the maximization of ATP or the enzymatic efficiency. Under the above consideration, the observed flux distribution of the system should be close to the flux distribution obtained by the optimization of the proper objective function and the proper set of constraints that describe and characterize the system and its environmental conditions.

Using an objective function, the constraint problem is reduced to an optimization task (Box 3.1). If the optimization function is linear with respect to the involved fluxes then the optimization problem is a linear programming problem and can be solved efficiently and exactly providing the flux distribution of the system. In prokaryotes, the most commonly used objective function concerns the maximization of the biomass production (a linear objective function), which is based on the hypothesis that unicellular organisms have evolved towards maximal growth performance. The flux through all reactions that drain biomass constituents, which include all the biosynthetic precursors and cofactors, is maximized.

The optimal function of the metabolic cell system at steady state as constrained by the capacity bounds is mathematically described in Box 3.1.

maximize	$Z = \vec{c}\,\vec{v}, \quad Z$: objective function, $\vec{c}$: known coefficient vector	Box 3.1
subject to	$S\vec{v} = \vec{0}, \quad S$: stoichiometric matrix	
	$\vec{lb} \le \vec{v} \le \vec{ub}, \quad \vec{lb}$: lower bound, $\vec{ub}$: upper bound	

Flux balance analysis (FBA) models search the feasible space (Fig. 3.6) and estimate the *optimal* flux distribution of the entire biochemical reacting system providing a quantitative description of the system and its metabolic capabilities when the intracellular fluxes are in balance. It might however be the case that the FBA problem gives more than one equivalent optimal solutions (*degeneracy*). An example of having (infinite) multiple optima is shown in Fig. 3.6 (d) where the level set corresponding to the optima is parallel to one of the facets of the flux cone. Issues related with the selection of the objective, the optimal solution space and the constraints of the system are discusses within the following sections as they determine the quality of the predictions of the model.

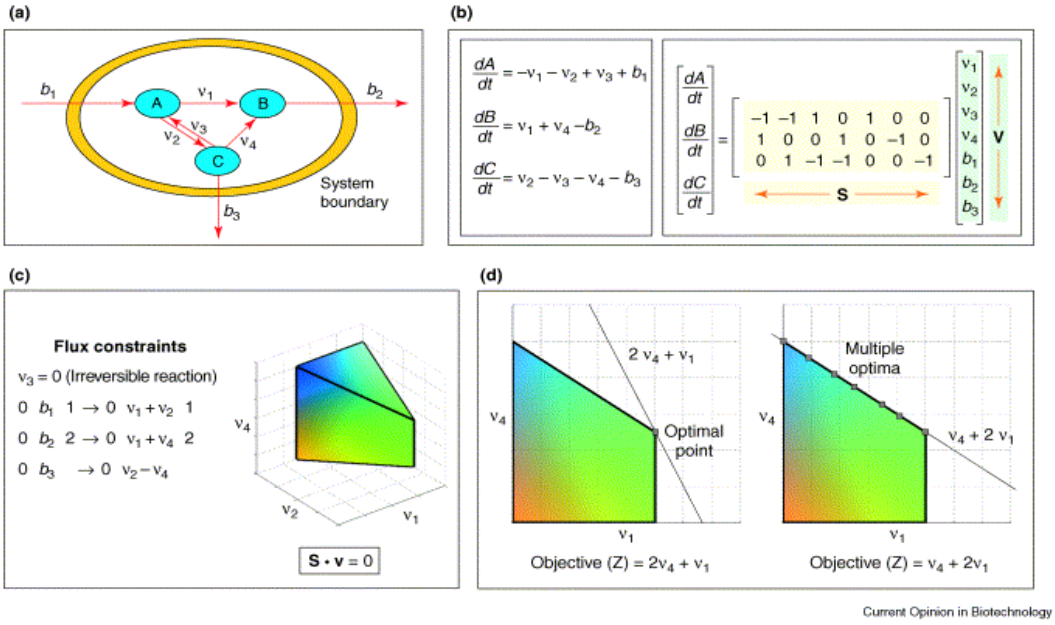


Figure 3.6 Flux Balance Analysis method applied in a simplified metabolic system. (a) A simplified chemically reacting system is shown. The external fluxes entering (b_1) and leaving (b_2 and b_3) the system through the system boundary are shown. (b) The system of the coupled differential equation can be described through the stoichiometric matrix *S*. (c) The steady state assumption and specific constraints bound the solution space (flux cone). (d) An objective function is applied to determine the optimum flux distribution of the system under the constraints. Single or multiple optima might be found. The picture is copied from [41].

3.4.4 Integrated Flux Balance Analysis

The solution obtained by FBA is as good as the assumption and constraints used to determine it. It is possible the feasible space that determines the real capabilities of the system and the feasible space determined by the assumed constraints to differ so that the observed solution does not coincide with the predicted optimal flux distribution. For the model to better represent the real biological systems and accurately predict the experimental evidences, further improvements concerning both the constraints and the optimization functions have been suggested since the method originally proposed. The thermodynamic constraints were developed to take into account the intracellular and environmental conditions of the system [179]. Flux limitations in respect to the way the system responses and adapts to genetic perturbations were also investigated [180]. Physical and spatial constraints resulting from high concentrations of macromolecules within the cytoplasm [181] have been considered as well. Furthermore, the incorporation of the temporal constraints to the model was also studied. The stoichiometry constraint implies that all the metabolic reactions present in an organism do actually participate and all gene products are available to contribute to the certain task-optimal solution. However, the system may utilize different metabolic pathways to respond to the various conditions as also throughout its temporal evolution. Consequently, an 'active' subset among all the possible metabolic reactions, actually plays role in the system dynamics [46, 172, 182, 183].

The method has proved successfully in analyzing the metabolic capabilities of several organisms, including its ability to predict deletion phenotypes, to determine the relative flux values of the metabolic reactions, to identify alternate optimal growth states [48, 50, 57], to guide the iterative refinement of model and to validate the metabolic network reconstruction [165, 183-185], to identify a group of reactions being active under all environmental conditions [186]. In *silico* predictions that concern growth rates, uptake rates as well as secretion rates of the *E. coli* metabolic network have proved to be consistent with experimental data under certain conditions [187].

In the following the optimality assumption, the uniqueness of the optimum solution as well as further constraints that may play key role in the determination of the optimum solution and the functioning of the cell system are discussed.

3.4.4.1 Optimality

Can all cell behaviors be described by a global objective? Does the organism obey in the same optimality principles in all environmental signals it receives? Doesn't the objective function depend on the applied environmental conditions and the functionality, specialization or even evolutionary short or long history of a certain cell?

The purpose of living systems for proliferation is evident. Towards the purpose of life evolution has impact on living systems to optimize their growth performance. This is why optimal growth has been widely used as the cell's objective. However, the organism will evoke all possible strategies if necessary in order to survive in a certain environmental condition. Thus, even in bacteria the assumption of best growth performance might not be a proper assumption under environmental stresses where the organism hasn't met before throughout its evolutionary history [188].

Furthermore, best performance from the cell's perspective might not simply mean maximization of biomass but best tradeoff between enzymatic cost and biomass production or enzymatic cost and ATP yield. Experimental flux data have been used as a ground truth in a decision problem for the determination of the most appropriate, among a group of possibilities, objective function which better describes the system under study and the data it produces [55]. For the six different growth conditions that *Schuetz et al* tested in that study, the following conclusion, which we consider important to mention, came up:

a. Under nutrient scarcity in *chemostat* cultures the operational state appears to have evolved under the objective to maximize either the overall ATP or the biomass yield.
b. For unlimited growth on glucose in oxygen or nitrate respiring batch cultures, the best optimization criterion is by far the maximization of ATP yield per unit of flux, a function that tries to trade of the overall ATP and the overall intracellular fluxes leading to a more economical allocation of resources.
c. In non-respiring batch cultures, the system operates in a low degree of freedom and all objective functions produce equally well predictions.

The above observations indicate that bacterial cells operate to maximize ATP yield but with respect to the enzymatic cost. However, under nutrient scarcity the enzymatic efficiency is relaxed. If a culture of bacteria grows in a dynamic environment of excess of nutrients, finite though, then isn't it reasonable to assume that in a period of time in which the nutrients become limited the organism should shift its optimization criteria to survive?

Apart from environmental signals, genetic perturbations affect cell's strategies as well. How does the cell actually respond to gene knockouts? The first hypothesis was that the cell under a genetically engineered knockout condition proceeds to a minimum possible redistribution of its fluxes. The method [189] is known as minimization of metabolic adjustment (MOMA) and employs quadratic linear programming to identify flux redistribution closest to the wild-type flux distribution. Experiments, however, have shown that under stressful conditions the organism initially responds with rapid and significant alterations in global gene expression patterns and eventually adapts to the new condition reaching a new steady state close to the one that has shown before perturbation. The regulatory on/off minimization (ROOM) method [180] based on the above observations, attempts to predict metabolic states after knockouts by minimizing the number of significant flux changes with respect to the wild type.

An interesting interplay between flux plasticity and network plasticity is evident and indicates the cell strategies that the organism evokes in order to efficiently respond to stressful conditions. The methods though are brute-force in a way that they do not attempt to mimic the regulatory strategies rather try to approach the global, phenotypic consequence of the actual response.

The determination of a proper objective function to represent the organism's real strategies remains a challenge. Furthermore, whereas the maximization of biomass production seems to be a plausible assumption for rapidly growing and replicating primitive cells such as bacteria for most of the cases, the flux distribution of cellular metabolism in the more complex eukaryotic cells is governed by a variety of cellular functions that have to be accomplished in parallel. Higher organisms including tissues and organs are not simple and should be handled in a more sophisticated way.

In parallel to understanding the fundamental biological principles and cell objectives, biotechnology is also interested in the formation of objective functions that satisfy an engineering goal such as the overproduction of a certain metabolite [58]. Although really interesting the identification of strains for obtaining improved, desirable phenotypes is beyond the scope of this discussion.

3.4.4.2 Optimal Solution Space

The optimal flux distribution might not be unique for a certain environment. Researchers have considered the possibility of having multiple optimal solutions in the flux space [190-193], a problem also known as degeneracy. Specifically this means that for any given optimal flux distribution there might be alternate solutions that generate the same objective value via different flux patterns. Those flux distribution patterns comprise the optimal solution space. Thus, in order to extract informative conclusions from the flux distribution of a given network that operates at an optimum mode, all possible solutions should be taken into account. It is unknown whether the whole optimal solution space is biologically meaningful. Definitely there are constraints still missing in FBA models, so that the optimal solution space can be the result of lack of knowledge. On the other hand, biological systems are both robust and evolved and alternative strategies can also be reflected in the optimal solution space to support the inherent functional heterogeneity. Nevertheless, finding all the alternate optimal patterns is not a trivial task.

Flux Variability Analysis

Redundancy and robustness, evident in biological systems, supports the existence of multiple optimal flow channels from environment to biomass production. Flux Variability Analysis (FVA) has been used to determine the ranges of fluxes that lead the system to a certain optimal objective value determined by FBA under given constraints [190]. FVA does not identify all possible alternate optimal solutions, just the range of flux variability. Flux variability is highly dependent on environmental conditions and network composition. Under gene knockout experiments in which sufficient evolutionary pressure is absent the growth rate varies significantly if the flux variability among alternate optima is high [190].

The flux variability analysis as described by Mahadevan et al [190] consists of the following optimization problems. The first step concerns the already described linear programming optimization problem that determines the optimum flux distribution of a given metabolic network under certain constraints. Many different flux distributions are possible to result in the same optimum value of the objective function determining the optimum flux space for the system under study. In order to determine the range that each flux in the network can vary while the system continuous to operates in optimum mode a series of linear programming problems are held. The new objective is the maximum possible value of the certain flux and subsequently the minimum.

For each v_i

maximize/minimize v_i

subject to $S\vec{v} = \vec{0}$, S: stoichiometric matrix

$\overrightarrow{lb} \le \vec{v} \le \overrightarrow{ub}$, $\overrightarrow{lb}$: lower bound, $\overrightarrow{ub}$: upper bound

$Z = \vec{c}\,\vec{v} = Z_{optimum}$, Z : objective function

Box 3.2

The principle of minimal effort

The principle of minimum effort states that cells capable of fulfilling vital functions such as growth with minimum effort should have had a selective advantage. The principle of minimum effort has been formulated differently. Minimization of the total flux of the network [194] is one option. Finding the minimal (active) subnetwork has also been proposed [192]. Alternatively, minimize the absolute value of fluxes (1-norm) in the optimal solution has also been applied (Box 3.3). If alternate optima have been found from the FBA problem, a second

optimization problem (Box 3.3) can be applied to determine the solution that costs less with respect to the amount of enzymes used by the cell.

$$\begin{aligned} &\text{minimize} \quad \left\|\vec{v}\right\|_1 \\ &\text{subject to} \quad S\vec{v} = \vec{0}, \quad S: \text{stoichiometric matrix} \\ &\qquad \overrightarrow{lb} \le \vec{v} \le \overrightarrow{ub}, \quad \overrightarrow{lb}: \text{lower bound}, \overrightarrow{ub}: \text{upper bound} \\ &\qquad Z = \vec{c}\,\vec{v} = Z_{optimum}, \quad Z: \text{objective function} \end{aligned}$$

Box 3.3

3.4.4.3 Variant constraints

Reconsidering the optimization principle and investigating cell's real objectives is one direction towards the development of the first generation constraint-based modeling. The incorporation of further constraints to the model, which are present in the biological systems, reduce the space of its metabolic capabilities and better define cell's possible behaviors, is the second alternative.

The objectives depict the cell strategies which actually hide the complex network of sensory and regulatory proteins that is orchestrated at both transcriptional and posttranscriptional levels. This regulatory system reflects the time variant and condition dependent constraints of the system, determines the active metabolic pathways and defines their corresponding fluxes. In that sense, cellular constraints and objectives are interrelated. Shedding light to all those interactions that take place at all levels from genomic to enzymatic activation is also important and remains far from completeness.

Evolutionary pressures have forced bacteria to adapt to variant environmental conditions which are present in their natural environment. Considering the fact that a bacterium utilizes the available sources in an effective and optimum -with some respect- way to produce biomass for its own benefit there is a network where signal propagates from environment to the corresponding biomass products. Not all the set of the possible metabolic reactions participates in the accomplishment of the metabolic tasks. Alternative pathways from the source to the biomass products are available making the organism flexible to different environmental conditions. Furthermore, adjustments in gene activation influence the enzymatic activity and temporarily determine the active metabolic channels for a given environment. Which pathway-set of reactions does the organism eventually decide to follow? The answer depends on both the environmental conditions and the underlying regulatory network that affects metabolism.

Dynamic Flux Balance Analysis

Flux balance analysis has been developed to embody in its original framework dynamic phenomena that affect the metabolic capabilities of the system and where the classical FBA model is not capable of describing. The dynamic FBA framework can thus describe phenomena such as the *diauxic* growth, in which a microorganism sequentially (and not simultaneously) metabolizes two sugars that are provided for its growth. Furthermore, the dynamic FBA framework can provide information about the metabolite concentrations whereas it can allow in its framework the integration with kinetic models for cases where the kinetic parameters are known. A static [47, 195] and a dynamic optimization [46] approaches have been mainly proposed into that direction. The dynamic optimization involves optimization of the objective over the entire time period of interest providing the time profiles of the fluxes and metabolites of the system. The problem is transformed into a nonlinear programming problem, which in general is computationally intensive whereas the solution it provides is similar to the static optimization approach. In the following the static optimization approach as proposed by Varma and Palsson [47] is analytically described.

Dynamic Flux Balance Analysis: static optimization

To simulate dynamic phenomena [47], the whole time regime that represents the time of growth in cell populations is properly divided into time slots where the current availability of the resources initiates the boundaries of the uptake fluxes (capacity constraints). The size of the time interval determines the smallest time constant of the model and should be in agreement with the smallest relevant time constant of the system under study [196]. The dynamic flux balance analysis includes the effect of a slowly varying environment which describes batch and fed-batch cultures. Optimum operation of the system within each time interval is assumed for the system to effectively reach its goal of growth and development (greedy algorithm). Given the initial concentrations of the concentrations of the substrates and the biomass, their transient changes can thus be predicted and compared with real data.

Let b_0 be the initial biomass concentration (gram per liter), exC_0 be the vector of the initial concentration values (millimoles per liter) of the exchange molecules (substrates) and δt be the time step. The dynamics (Box 3.4) and constraints (Eq. 3.3) that describe a batch culture are the following.

$$\frac{db}{dt} = \mu \cdot b \quad \rightarrow \quad b[t] = b[t-\delta t]e^{\mu \cdot \delta t}$$

$$\frac{\partial exC}{\partial t} = v_{ex} \cdot b \quad \rightarrow \quad exC[t+\delta t] = exC[t] - v_{ex}\frac{b[t]}{\mu}(1-e^{\mu \cdot \delta t})$$

Box 3.4

The growth rate μ and the fluxes v of the system including the exchange fluxes v_{ex}, which correspond to the substrates, are time-variant variables and are determined by the FBA at each time step. The FBA optimizes the growth rate μ (objective) under the substrate availability constraints (Eq. 3.3), which bound the uptake fluxes (capacity constraints) of the system in order to conserve mass. Thus, the current concentrations of the substrates (exC) are scaled by the amount of the current biomass b (resource partitioning) to shape the boundaries of the uptake fluxes. The dynamic Flux Balance Analysis is also described in Algorithm I.

$$v_{ex}^{bound} = \frac{exC}{b \cdot \delta t} \qquad \text{Equation 3.3}$$

It is important to mention that multiple optimum solutions [190] might exist for the same instance of the problem. Furthermore, when a choice that represents a certain phenotype is made the environment is affected and the optimum solution space of the next time step is shaped accordingly. In simulations however, only one optimum solution is projected to the next time interval (Fig. 3.7). The range of values within each flux in the network can vary, while the system continuous to operate in optimum mode can be investigated by applying a series of linear programming problems. This method is known as flux variability analysis [190] and described previously as well as other approaches that attempt to explore the optimum solution space (3.4.4.2).

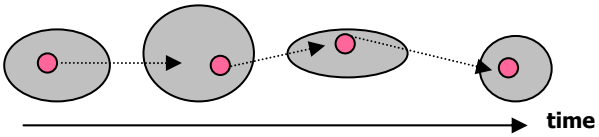


Figure 3.7 The grey areas represent the feasible flux spaces for different time intervals where the optimum flux distributions of the metabolic system (depicted with pink color) operating at steady state should lie. When a choice that represents a certain phenotype is made among other alternatives the environment that determines the constraints of the next time interval is affected accordingly and thus the sequential optimum solution space. In dynamic simulations only one optimum path is usually selected.

Algorithm I

For each time interval $(t-\delta t, t]$

maximize $\mu = \vec{c} \cdot \vec{v}$, μ: growth rate, $\vec{v}$: flux vector

subject to $S \cdot \vec{v} = \vec{0}$, S: stoichiometric matrix

$\overrightarrow{lb} \le \vec{v} \le \overrightarrow{ub}$, $\overrightarrow{lb}$: lower bound, $\overrightarrow{ub}$: upper bound

$$b[t] = b[t-\delta t] \cdot e^{\mu \delta t}$$

$$\overrightarrow{exC}[t+\delta t] = \overrightarrow{excC}[t] - \overrightarrow{v_{ex}} \cdot \frac{b[t]}{\mu}\left(1 - e^{\mu \delta t}\right)$$

$$\overrightarrow{ub_{ex}} = \overrightarrow{exC} / (b \cdot \delta t)$$

end

where

$\overrightarrow{v_{ex}}$ is the flux of the exchange reactions

$\overrightarrow{ub_{ex}}$ is the ub of the exchange reactions

$b[0] = b_0$, $\overrightarrow{exC}[0] = \overrightarrow{exC_0}$

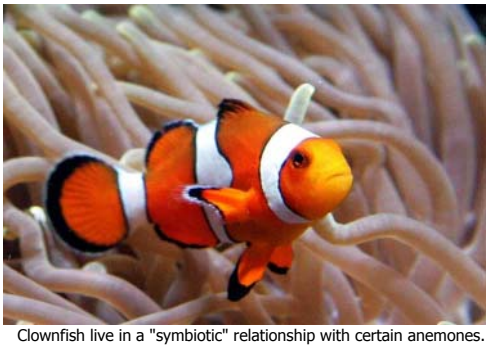
Regulated Flux Balance Analysis

The gene regulatory system among other responsibilities is a server that works to offer to the metabolic network the proper enzymes needed at certain time under the current environmental conditions. Gene regulation comprises the cardinal mechanism that determines which and when metabolic genes will be available to contribute to the metabolic demands. However, gene regulatory networks are only partially known even for well-studied organism such as *E. coli* [197, 198]. Furthermore, there is an information gap between gene activities and the corresponding enzymatic activities that actually determine the active pathways in the metabolic network. Post-transcriptional mechanisms and synergetic phenomena do actually take place that turn the enzyme activity to unpredictable from gene expression data only. Therefore, the expression of a metabolic gene doesn't necessarily mean enzymatic activity. The condition is necessary but not sufficient. However, if a gene is not expressed the corresponding protein and its related activity would be absent [199].

The activation of metabolic genes is mainly regulated by the set of their related *TFs*, which are affected in turn by both the extra-cellular (environment) and the intra-cellular metabolites (metabolic products). Regulatory rules are applied between extra/intra cellular metabolites

and *TFs* as well as between *TFs* and metabolic genes in order to mimic the regulatory strategies of the cell and the temporal evolution of gene–gene product interactions [183, 200].

The regulated flux balance analysis model is based on regulatory rules to dynamically produce a set of metabolic genes which eventually determine the enzymatic activities; thus it provides a temporal trace of the cell's metabolic capabilities. Usually the incorporation of gene regulation in metabolism begins with the application of Boolean rules that bridge the internal system with its environment (external conditions such as presence/absence of extra cellular molecules) as well as rules that determine the enzymatic activity/inactivity in a given environment [172]. Example of rules: a gene transcription actually does take place if a gene is active and an end product metabolite is not present. A reaction takes place if the chemical products are available and the proper enzyme present. How the presence of the next time step enzymes is determined? Enzymes present in the first time step are on until a degradation time passes or an end product metabolite suppresses their expression. Whether such kind of regulatory phenomena can be represented by Boolean logic remains an open question.



Clownfish live in a "symbiotic" relationship with certain anemones.

4. Cross-feeding and Efficiency: Reconstruction-Modeling

4.1 Introduction

Cross-feeding is the exchange of nutrients among individuals arising from their metabolism. Cross-feeding interactions can play an important role in the evolution of diversity within populations and the establishment of cooperative strategies. Interacting cell communities, unlike single cells, dynamically shape the environment and the fitness landscape allowing new metabolic capabilities that were not expected before to emerge. Polymorphism maintained by cross-feeding has been also observed to emerge in long-term evolution experiments of *E. coli* populations initiated from a single clone, which grow on a single-limited resource (see section 1.5).

A mathematical representation and a computational framework that allows the efficient identification, exploration and comparative analysis of potential cross-feeding interactions in a pool of strains, in various growth conditions are introduced in this work. A graph representation is reconstructed to express the potential cross-feeding interactions.

Furthermore, a novel, growth model capable of incorporating metabolic interactions between self-interested members of a heterogeneous population while it describes cells at genome-scale, is proposed. The model is called multi-competitor growth model and it is based on the dynamic Flux Balance Analysis model (section 3.4.4.3).

Under the plausible hypothesis that the best use of resources enhances the likelihood of survival subject to ecological and evolutionary constraints, communities composed of self-interested strains of efficient, improved growth performances, capable of better utilizing the available resources comprise an important aspect of this study and are thoroughly investigated. Several definitions with respect to the growth performance and the efficiency of a heterogeneous population are given here. It is mathematically proved (Appendix) that in a simple and spatially homogeneous environment where by-production is not allowed to disturb the medium, competition for the primal source alone cannot lead a heterogeneous population to group benefit supporting the hypothesis that other sources of heterogeneity such as by-production might play a critical role in growth efficiency.

4.2 Metabolic Diversity Graph reconstruction

The *diversity* graph $G = (V, E, w)$ consists of a set of nodes V, a set of edges $E \subset VxV$ and a function $w: E \to [0,1]$, which assigns a real number in $[0,1]$ to every edge of the graph encoding the level of difference between two nodes. Under this definition, zero weight reflects identical nodes and maximum weight reflects highest difference.

Depending on the problem and the properties encoded in the nodes, the diversity graphs can be reconstructed in order to identify the potential to exchange products, ideas, or information within a group.

Given a pool of genetically different strains, a graph representation is introduced to quantitatively reflect the metabolic variability of these strains allowing reduction of the search space of potential strain communities. The hypothesis is that when bacterial cells are growing on a single-limiting resource, metabolic interactions that involve the exchange of intermediate metabolic products might occur if there are differences in the metabolic capabilities of the members of the bacterial population with respect to by-production. The metabolic products disturb the homogeneity of the growth environment, comprise the mean of exchange in an interacting population and provide the raw material of interesting phenotypes to emerge. Therefore, the edges of the graph are defined in a way to quantitatively describe differences in the metabolic capabilities of the strains with respect to by-production and because of that that graph is named *diversity graph*. Considering that the strains are derived from specific genetic perturbations applied on the cell, the larger the number of the genetic perturbations, which affect a cell's phenotype with respect to by-production in a given environment is, the higher the metabolic variability and consequently the ecological opportunities and the potential of the system to generate polymorphism is expected to be.

The nodes of the metabolically diversity graph correspond to viable strains. Among all possible interactions between the genetically different strains, those that provide each other with different metabolic capabilities with respect to by-production comprise the edges of the diversity graph. The metabolic capabilities of the cells vary according to the growth conditions, therefore growth on different nutrient sources results in different diversity graphs. The metabolic capabilities of each strain in a specific growth environment are estimated by the dynamic Flux Balance Analysis method (section 3.4.4.3).

The dynamic Flux Balance Analysis method takes into account the genome-scale metabolic reconstruction of the cell, correlates the genomic information to the metabolic activity and enables the prediction of the relative flux values of the metabolic reactions, which correspond to the optimum flux distribution of the network under the given constraints [54]. Beside the growth phenotype predictions, the model can also provide information about the metabolic by-product secretion of the cell, which has been shown to be in consistence with experimental data for specific environmental conditions [47, 57].

Node property: metabolic capabilities

The dynamic growth of each strain on a given single-limited resource is simulated. Mutants that cannot grow in the certain environmental condition (lethal mutations) are excluded from the graph. Given a specified initial quantity of food, the time profiles of the growth rate, the concentrations of the exchange molecules and the biomass concentration are estimated. A definition for the metabolic capabilities of each strain under the perspective of its potential to metabolically interact with another strain through exchange of nutrients is essential.

It is assumed that the metabolic capabilities of a strain can be reflected by the maximal amount of the byproducts that the strain is capable of providing to the pool of public goods. The metabolic capabilities of a strain are thus quantified based on the amount of the

metabolites that are by-produced during the metabolism. A vector consisting of the maximum amounts of each byproduct is reconstructed for each strain. This feature vector comprises an important metabolic blueprint for the potential interactions of the strain. For the feature vectors to be comparable with each other, the intersection set of the byproducts of all strains is defined. If a specific strain is not capable of producing a specific metabolite appeared in the intersection set then this entrance in its vector equals zero, otherwise it takes the maximum amount of the metabolite produced by the strain.

Edge property: metabolic difference

The maximal metabolic difference of two strains can be expressed with their maximal concentration difference over all byproducts. Specifically, over all byproducts (s), the maximal absolute relative concentration difference of the peaks is used as a weight (w_{ij}) for the interaction between two nodes (i and j), leading to the reconstruction of a naturally weighted graph, where the peaks ($_{\max}C_i^s$) are defined as the maximum value of the concentration of a substrate over time. When one strain provides a novel byproduct to another strain, the weight equals to 1 under the above definition. Depending on the information we want to infer from these graphs, a threshold can be set that reflects the level above which the concentration differences are considered important.

$$w_{ij} = \max_s \left(\frac{\left| {}_{\max}C_i^s - {}_{\max}C_j^s \right|}{\max({}_{\max}C_i^s, {}_{\max}C_j^s)} \right), \quad \text{where } {}_{\max}C_i^s = \max\left(C_i^s(t)\right) \qquad \text{Equation 4.1}$$

The relative concentration differences of the by-products that determine the edges of the diversity graph actually express the relative by-production efficiency between the mutants that is defined as the ratio of the by-production rate to the uptake rate of the main source (Fig. 4.1). Thus, the diversity graph is independent of the initial concentration of the primal source that is provided to the system for growth (Proof in 4.2.1). This observation is of particular importance since it allows a unique graph representation of the metabolic properties of the mutants for a given growth condition that depends only on the applied constraints and assumptions of the metabolic model. It further implies that the diversity graph can be also estimated by the simple Flux balance Analysis model, which accelerates the reconstruction of the diversity graphs. Alternative edge definitions are presented in 4.2.2. It is important to mention that the assumptions implied on the definitions of edges determine the search space where potential interacting communities can be found and shape the predictions of the model.

4.2.1 Structural robustness

The question here is whether the reconstruction of the diversity graph depends on the initial amount of the main source or remains robust when this quantity changes. The robustness of the structure can be determined by the way the edge weights are actually affected under changes in the initial concentration of the main source.

The concentration value of an exchange molecule at a time t depends on the uptake for consumption (or intake for production) rate u of the specific molecule, the growth rate μ, the initial biomass b_0 and the initial concentration value C_{t_0} of the molecule according to Eq. 4.2. A byproduct p reaches its maximum concentration value at the time t_{exh} (Eq. 4.4) where the main source m is getting exhausted (Eq. 4.3).

Under the hypothesis that during the growth phase, which corresponds to the metabolism of the main source, the growth rate and the uptake (or intake) rate, remain constant the amount of the produced nutrient is proportional to the initial amount of the main source (Eq. 4.5). The hypothesis of constant flux rates is actually true as long as sufficiently small time intervals can be considered. In that case, the relative concentration difference of a byproduct that determines the edge weight according to Eq. 4.1 is proved to be independent on the initial amount of the main source. Therefore, the diversity graph representation remains unaffected under changes in the initial amount of the main source, whereas the edge weights express the by-production yield (Eq. 4.6).

$$C_t = C_{t_0} - \frac{u}{\mu} b_0 (1 - e^{\mu t}) \qquad \text{Equation 4.2}$$

Main source:

$$C^m_{t_{exh}} = 0,\ [1] \Rightarrow C^m_{t_0} = \frac{u^m}{\mu} b_0 (1 - e^{\mu t_{exh}}) \qquad \text{Equation 4.3}$$

By-production:

$$C^p_{t_0} = 0,\ [1] \Rightarrow C^p_{t_{exh}} = \frac{u^p}{\mu} b_0 (1 - e^{\mu t_{exh}}) \qquad \text{Equation 4.4}$$

$$[\text{Eq. 4.3}], [\text{Eq. 4.4}] \Rightarrow C^p_{t_{exh}} = \frac{u^p}{u^m} C^m_{t_0} \qquad \text{Equation 4.5}$$

Edge weight, w_{ij} between two strains i and j with respect to product p:

$$w_{ij} = \frac{\left| C^p_{t_{exh}}(i) - C^p_{t_{exh}}(j) \right|}{\max\{C^p_{t_{exh}}(i), C^p_{t_{exh}}(j)\}} = \frac{\left| \frac{u^p(i)}{u^m(i)} - \frac{u^p(j)}{u^m(j)} \right|}{\max\{\frac{u^p(i)}{u^m(i)}, \frac{u^p(j)}{u^m(j)}\}} \qquad \text{Equation 4.6}$$

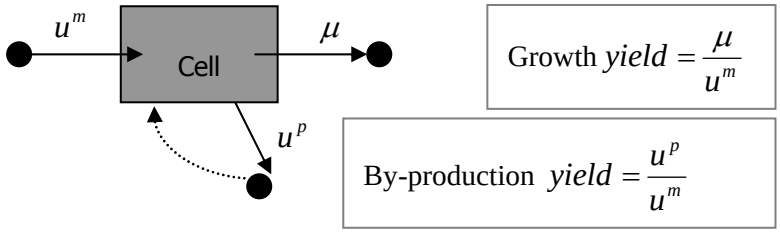


Figure 4.1 A simplified diagram showing the main mechanisms of biomass production and by-production during the metabolism of the main source.

In simulations however time is discretized into finite time intervals. The consequence of this discretization is that the flux capacity constraints of the system are affected when the main source is shifted from abundance to scarcity. Thus, the constraints of the system close to the time where the main source is getting exhausted may add a small inaccuracy in the simulations that may alter the edge weights of the diversity graph.

4.2.2 Alternative edge definitions

Additional properties that may determine the edges of the diversity graphs in different ways concern the ability of the involved strains to consume the byproducts.

Constrained: The metabolic difference is taken only upon byproducts that at least one of the two potential participants is capable of consuming. Novel byproducts for one of the two potential participants are constraint-free. This condition on the edges of the graph implies that any two metabolically different strains also have the potential to exchange nutrients (metabolically interact).

Unconstrained: The 'consumability' constraint is relaxed. The relaxed condition implies that two strains are allowed to both provide a certain nutrient to the public pool without consuming it, where however, a third strain of different metabolic capabilities of both the two strains can exploit. Thus, the metabolic difference between two strains might be important in a divergent community, even if it does not lead to direct exchange of nutrients.

Super-unconstrained: This class includes cases where independently of the value of the metabolic difference, strictly one of the two participants is not capable of consuming a specific by-product. These edges express provider-consumer relations of known metabolites and when inserted in the graph take weight values equal to 1.

4.2.3 Alternative graph representations

Binary: Apart from the weighted representation, a binary representation is also employed. Depending on the properties to be inferred from these graphs, a threshold can be set to reflect the level above which the concentration differences are considered important, leading to a binary representation of the diversity graph.

4.2.4 Graph compression

Naïve: All strains with the exactly same metabolic capabilities under a given initial environmental condition are grouped together. From this partitioning of the nodes to classes, a representative node of each class is arbitrarily chosen to form the compressed diversity graph. Each class actually contains all the single gene deletions that have exactly the same effect on cell functioning under a certain condition.

Structural: The structural compression maps all the structurally identical nodes of same connectivity onto a super node. This compression can be used to fasten the clique identification problem that is discussed later, whereas it can also allow the graph to be nicely visualized by highlighting the nodes and interactions that actually produce the metabolic diversity of the system.

Functional: In order to reduce the number of growth simulations the diversity graph is (functionally) compressed in the number of nodes. The functional compression utilizes the general robustness of the metabolism to genetic perturbations and forms a class of those strains that have growth properties similar to the wild-type cell. The maximum norm and a cutoff of 10^{-3} are used as a measure of similarity. Under heterogeneous population growth these gene deletions representing different constraints in the metabolic network might not have the same phenotypic effect when novel substrates appear in the environment.

4.3 Finding Strain Communities

The composition of a metabolically interacting strain community can be assumed to consist of individuals with the potential to differently shape the given environment and provide each other products of their metabolism. If a strain v is metabolically similar to a strain k, then either v or k can be a member of the strain community but not both. The addition of genetically different but metabolically similar mutants in a strain community does not provide

any new experience in the growth medium and phenotypically is as one strain, which appears in the community in a twofold initial population ratio.

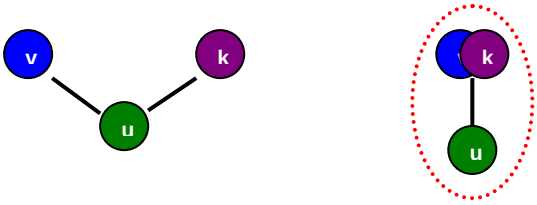


Figure 4.2 If a strain v is metabolically similar to a strain k, then either v or k can be a member of a strain community. Both are metabolically redundant.

Clique: strain community

The property of difference ('$\neq$') as defined for the metabolically diversity graphs is non-transitive: the knowledge that the nodes $v \neq u$ and $u \neq k$ does not provide any information about the relation between the nodes v and k, only the existence or non existence of an edge between them ensures their relation.

To ensure compositions of strains with different metabolic capabilities each strain must be different from all the rest strains. Therefore, compositions of strains with different metabolic capabilities correspond to cliques (complete sub-graphs in which all the nodes are connected with each other - see 2.3.4) in a diversity graph.

The number of cliques can grow exponentially with every node added in the graph and finding all cliques is expensive in general. However, because of the genetic robustness and redundancy, viable mutants of novel metabolic capabilities are in general few allowing significant compression to the number of nodes (structural compression). Cliques are found in the diversity graphs using their binary representations.

Maximum clique: upper bound of metabolic variability

The maximum clique is a clique of the largest possible size in a given graph. The maximum clique size that can be found in the diversity graph reflects the actual number of the different metabolic patterns, which also determines the upper bound of the potential diversity that can emerge in the system under the given growth condition and within the search space that is defined by the allowable genetic perturbations.

4.4 Comparative Analysis of Metabolic Diversity

In this work, biologically interesting graph properties such as the network centrality, the assortativity and the clustering coefficient (section 2.3) are explored in order to characterize and compare the different diversity graphs that correspond to the different growth conditions.

The strength (or degree) centrality is a measure of the importance of a node in a graph. In the diversity graphs, a highly central nodes indicates a strain of considerably different (unique) metabolic capabilities with respect to by-production from the rest strains of the graph whereas a highly non-central node correspond to a strain of similar (common) metabolic capabilities with each other.

The main questions that are addressed concern the way the inherent metabolic redundancy is reflected on the graphs and quantified in their properties as well as the growth conditions in which the organism is more vulnerable to genetic perturbation increasing its probability to develop cross-feeding interactions and become polymorphic.

4.5 Multi-competitor Growth Model

Under the constraint-based framework (described in section 3.4.2), the cells grow metabolizing the nutrients provided in their environment under a variety of physiochemical constraints and the assumption that they have been evolved towards best exploitation of the given environment in order to maximize their growth rate or another cellular objective. The genome-scale metabolic reconstruction of the cell is utilized in the model in order to bridge the genotype with the phenotype and predict cellular behaviors.

In the static optimization approach (section 3.4.4.3) of the dynamic metabolic simulations cells are assumed to work greedy in the dynamic environment – optimizing for the best choice at each time step. It is reasonable to expect conditions where the local optimum does not imply and thus does not guarantee global optimum (end-point biomass optimization). Long-term efficient strategies which increase the amount of essential secondary metabolites – byproducts of the metabolism might prove beneficial with respect to the final biomass concentration, which measures the growth performance of the population. However, the greedy approach simulates adequately a competitive life (competition for food), allows adaptations to a dynamic fitness landscape and has proved successful in analyzing the metabolic capabilities of several organisms, including its ability to predict deletion phenotypes and determine the relative flux values of the metabolic reactions [47, 57].

Whereas specific and simplified growth and metabolic models have been proposed elsewhere (see section 1.9) to describe cross-feeding interactions and stable coexistences of different genotypes or species, multi-species relationships and multi-cellular communities haven't been developed thus far under the constraint-based framework which accommodates the genome-scale metabolic reconstruction of the species. Instead, in the metabolic simulations, all cells in the population follow the same metabolic and regulatory program.

This work extends the dynamic growth simulation approach from monocultures to polymorphic populations based on the constraint-based principles. A variety of cells can co-grow in a common environment and in accordance to their metabolic capabilities, they metabolize the available nutrients, produce secondary metabolites, and compete for the (primal and secondary) resources in order to best exploit the common environment for their own benefit, which is assumed to be the maximization of their growth. Their growth shapes the environment and allows the development of cross-feeding interactions through their products of metabolism. The static-greedy optimization approach is accommodated in the proposed multi-competitor growth model.

The dynamics and constraints that describe a shared, well-mixed batch culture consisting of different cells during the exponential and early stationary stage where the cellular death rate can be considered negligible are given by the equations shown in Box 4.1 and Eq. 4.6 respectively.

$$\frac{db_i}{dt} = \mu_i \cdot b_i \quad \rightarrow \quad b_i[t] = b_i[t-\delta t]e^{\mu_i \cdot \delta t}$$

$$\frac{\partial exC}{\partial t} = v_{ex}^{i} \cdot b_i \quad \rightarrow \quad exC[t+\delta t] = exC[t] - \sum_i v_{ex}^{i} \frac{b_i[t]}{\mu_i}(1-e^{\mu_i \cdot \delta t})$$

Box 4.1

The initial biomass concentration (b_0^i) of each competitor, which determines its initial relative frequency in the population, is defined. The initial concentration values of the exchange molecules (substrates) and the time step δt , which represents the temporal resolution of the system are also determined a *priori*. The growth rate μ_i and the fluxes v^i including the

exchange fluxes v_{ex}^{i} , which correspond to the exchange molecules with the environment, are time-variant variables and are determined by the FBA separately for each competitor at each time step. Thus, the number of FBA problems that have to be solved at each time step equals the number of the competitors. Each FBA problem optimizes the growth rate μ_i (or another selected objective) of each competitor respecting the availability of the substrates in the common environment, which bound the uptake fluxes of the system. The capacity constraints represent the conservation of mass with respect to the substrates and in general can be defined differently for each competitor. In this work, all cells are assumed to sense the same bounds (Eq. 4.6). This assumption benefits the larger populations, when the resources that shape the corresponding bounds become sparse in the medium. Alternatively a variable time step can be used.

$$v_{ex}^{bound} = \frac{exC}{\sum_i b_i \cdot \delta t}\ ,\ \forall i \qquad \text{Equation 4.6}$$

The static-greedy optimization approach of the dynamic growth of a multi-competitor system is analytically described in Algorithm II. The simulation terminates when none of the competitors can grow further in the shaped medium, which usually corresponds to the phase of nutrient depletion. A shared, well-mixed population (spatially homogeneous) environment is assumed. When the number of different cells-competitors is one, Algorithm II is reduced to Algorithm I. The competitors can represent different strains or species in which the genome-scale metabolic network is available.

Algorithm II

For each time interval $(t-\delta t, t]$

  For each competitor $i = 1:M$

    maximize μ_i

    subject to $S_i \overrightarrow{v^i} = \vec{0}$, S_i: stoichiometric matrix

    $\overrightarrow{lb_i} \le \overrightarrow{v^i} \le \overrightarrow{ub_i}$, $\overrightarrow{lb_i}$: lower bound, $\overrightarrow{ub_i}$: upper bound

    $b_i[t] = b_i[t-\delta t]e^{\mu_i \delta t}$

  end

  $\overrightarrow{exC}[t+\delta t] = \overrightarrow{exC}[t] - \sum_{i=1}^{M} \overrightarrow{v_{ex}^i} \cdot \frac{b_i}{\mu_i}\left(1 - e^{\mu_i \delta t}\right)$

  $b = \sum_{i=1}^{M} b_i$

  $\overrightarrow{v_{ex}^{i,bound}} = \overrightarrow{v_{ex}^{bound}} = \overrightarrow{exC}/(b \cdot \delta t)$

end

where

$\overrightarrow{v_{ex}^i}$ is the flux vector of the exchange reactions for the competitor *i*

$\overrightarrow{v_{ex}^{bound}}$ is the *ub* of the exchange reactions for each competitor *i*

4.5.1 Growth Efficiency

Efficiency concerns the ability of a system to maximize its growth performance (output) given a certain limited amount of resources (input). In the metabolic simulations of this work the performance of a cell population is measured with respect to the maximum (endpoint) total biomass concentration that the system is capable to produce in a given environment of a certain amount of nutrients.

In cases of systems that consist of different monoclonal (homogeneous) cell populations, the growth performances are compared with each other under the same initial conditions to determine the system with superior performance. In heterogeneous populations, the definition of a beneficial coexistence might have multiple aspects though. To quantitatively describe superior performance in these coexistences two definitions using different perspectives are given, the **absolute** and the **relative** benefit.

In a multi-competitor, heterogeneous system the growth performance of the group corresponds to the total biomass and is determined by the cumulative contribution of the growth performances of each of the competitors to the system. If the group performance of the heterogeneous cell population is superior to the performance of any (wild type or mutant) homogeneous population then the heterogeneous community under study is beneficial. We name this benefit **'absolute'**. However, the condition of 'any' can be relaxed, so that the growth performance of the heterogeneous population is compared to the homogeneous performances of all the members which constitute the community under study. In this case we call the benefit **'relative'**. The relaxed definition assures that when any of the component mutants appears it will not be capable to dominate the population, since coexistence is more beneficial. For bioengineering purposes the absolute benefit that corresponds to a simple way to produce more biomass might be more interesting, however, biodiversity might support the relative benefit as well. Furthermore, the growth performance of the wild type population can always be used as a benchmark.

4.5.2 Predictability of Growth performance

The growth performance of a bacterial community depends on the growth properties and metabolic capabilities of the competitors and the interactions between them. The characteristics of the competitors alone cannot explicitly determine the evolution of the interplay that takes place within a community and predict its growth performance. Cell communities, unlike single cells, dynamically shape the environment and the fitness landscape and give rise to new metabolic capabilities that were not expected before. However, if metabolic interactions do not take place in a community it is expected that relative benefit cannot be observed (main hypothesis) since the best performed and most efficient mutant would prefer to grow independently than co-grow with less efficient mutants.

It is worthwhile to further observe the multi-competitor system under the perspective of the interactions between the competitors in a way to give answers to questions such as whether the determination of the performance of a multi-competitor interplay can be related and reduced to the performances of the two-competitor systems involved or equivalently if it is possible that the performances of the pair-wise interactions among the competitors that appear in a bacterial community are sufficient to determine the performance of the whole community. The hypothesis thus to be validated is whether the performance of the community can be assumed to correspond to the sum of the growth performances of all the pair-wise interactions it consists of and how the true growth behavior actually depends on this mean performance of all the involved pair-wise interactions. Extending this argument the question can be rephrased to which are the minimum sized sub-cliques that a clique can be decomposed to so that knowledge of their performance is sufficient to predict the performance of the communities. This information actually defines the boundaries of the

propagation of novelty within communities. Any community of size above the sufficient has growth performance that can be predicted linearly.

Effective weights

Throughout the analysis for several different growth conditions it is observed that the performance of a clique linearly depends on the performances of the pair-wise interactions it consists of as long as pairs of mutants of unexplored or unexploited metabolic capabilities are not present to differently (non-linearly) contribute to the growth performance. However, for some cases this type of relations is present which increases the size of the minimum sub-clique that can sufficiently predict the growth performance of the communities to triplets. Mutant pairs which are not capable of interacting with each or fully exploiting the metabolites that they by-produce exist and play an important, novel role in larger communities. To validate the extend of the performance predictability (hypothesis) the concept of an effective weight is introduced in order to weight the interactions according to their potential contribution to the growth performance of the community instead of using their actual pair performance. The effective weights can be estimated by the simulation data coming from the cliques of size 3 (triplets). High divergence of the simulated growth performance of the triplets from the mean performance actually indicates the existence of unexplored metabolic capabilities in specific mutant pairs and the exploitation of these capabilities by larger communities (which here are the triplets). This approach is equivalent with the decomposition of the cliques of size greater than 3 to triplets. It is important to mention that consistent divergence from the mean performance (observed as groups of lines) and not a presence of a few outliers should come from edges between highly central nodes since these edges participate in many different communities in a given community size. For this reason the effective weights are searched and assigned to the edges of the highly central nodes only.

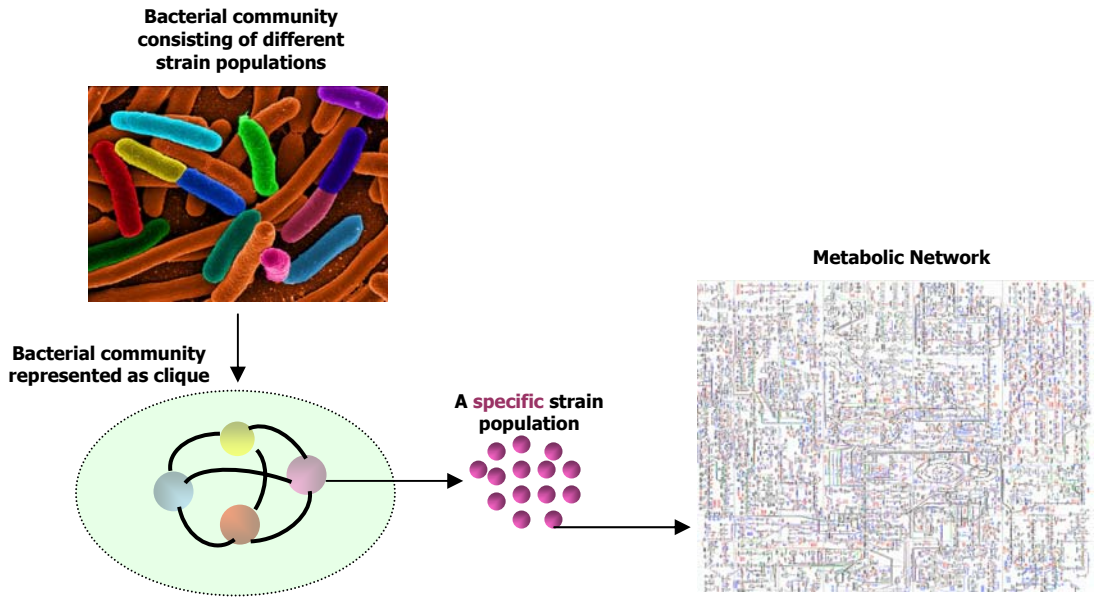
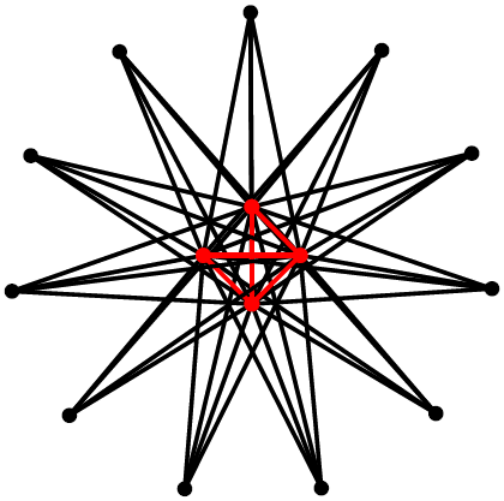


Figure 4.3 A bacterial community consisting of genetically and metabolically different strain populations corresponds to a clique in the diversity graph. Each strain exhibits metabolic capabilities determined by its metabolic network and the constraints of the dynamically shaped growth medium.



Metabolic Diversity Graph (see section 5.3). The structure reflects the evolved genetic robustness of metabolism. The Black and Red nodes correspond to strains of common (many) and unique (few) phenotypes respectively. The Red nodes form a highly clustered area and comprise the core of the potential strain communities.

5. Results

5.1 Introduction

The current work uses the bacterium *Escherichia coli,* as a case study. Genetic perturbations and specifically single, metabolic gene knockouts (KO) are applied to generate a pool of mutants among which potential cross-feeding interactions are examined.

In the Appendix of this work it is mathematically proved that when by-production utilization and cross-feeding are not allowed to take place, then the conservation of mass prohibits group benefit to emerge from two different populations that compete for a primal limited source independently of their initial frequency or the amount of the resource. An example of cross-feeding entails that under a condition which describes the initial ratio as well as the growth and metabolic characteristics of the competing populations, group benefit can emerge. The growth simulations are performed in order to reveal scenarios of by-production utilization where group benefit can be observed in communities comprised of single-gene knockout mutants given genome-scale descriptions of the cells and a constraint-based approach to describe their dynamic adaptations. Furthermore, cases where the performance of the group exceeds the growth performance of any single-growth mutant are investigated.

Simulations are performed using the genome-scale metabolic reconstruction of *E. coli* (*i*JR904) by Reed et al. [201], which includes 904 genes and 931 biochemical reactions. The gene-reaction association matrix is also provided. Each gene, encodes an enzyme that catalyzes one or more reactions and each reaction can be associated with one or more genes. The associations are modeled as logical relationships and include cases such as isozymes, multifunctional proteins and protein complexes. From the genes corresponding to the same set of reactions only one is considered in simulations because the rest have redundant phenotype under single-gene deletions. This filtering directly reduces the search space from 904 knockouts to 651. If a single gene is associated with multiple reactions, its deletion

affects all associated reactions. The upper and lower bounds for the reactions associated with the deleted gene are set to zero inhibiting the utilization of the specific metabolic reactions.

The applied genetic modifications concern evolved strains that are assumed to have optimal growth characteristics in any growth environment tested as this can be calculated by the Flux Balance Analysis framework. Maximization of the growth rate is used as the objective function of the optimization problem. A second optimization problem is also applied to minimize the enzymatic cost expressed by the absolute flux values under the constraint that the cell continuous operating at maximum growth rate (see section 3.4.2.2). The dynamic FBA framework proposed by Varma and Palsson [47] is used to simulate the population growth of each mutant in a dynamic environment. The proposed multi-competitor growth model is used to simulate the dynamic growth of the strain communities (competition experiments). Simulations are performed using the COBRA toolbox [202] and the modified version described in Algorithm II (see 4.5). The glpk solver [203] is used for solving the linear programming problems.

The initial bounds of the uptake rates are set in accordance to the work of Covert et al. [200]. The time resolution is set to 0.1 h (6 min). The initial biomass concentration is set to 0.003gDW/lt [200]. In the competition experiments the initial biomass concentration is equally distributed to the strains of the heterogeneous population under study unless it is differently specified.

Depending on the growth condition different pathways of the metabolic network are, in general, activated. Therefore, the same genetic perturbations can cause different metabolic capabilities with respect to by-production under different growth environments. In this work, 58 different carbon sources were examined since 10 of the 68 single-carbon source conditions defined in [200] did not show any growth in any single-gene deletions or the wild-type. Each time a different carbon source of initial concentration equal to 10mmol/lt is provided to the system for growth. Oxygen and ammonia, as well as other important inorganic metabolites are assumed to be in excess in the growth medium.

Given that all experiments are performed with the same initial amount of resources (system input), the end-point biomass concentration (system output) determines the efficiency of the population. The terms growth performance and end-point biomass concentration are used alternatively here.

1st set of experiments:

The first set of experiments includes an exhaustive study of the metabolic behaviors between the wild type (WT) cells and each mutant (KO^-) under the 58 different carbon sources. The aim of this study is the identification of mixed population (WT-KO^- pairs) with superior growth performances and the understanding of the metabolic interactions that may take place and lead the system to its superior performance. This first study subsequently inspired the diversity graph reconstructions and the exploration of the metabolic diversity that may arise in a bacterial cell population when single gene knockouts are allowed to perturb the system. The diversity graphs enable the efficient exploration of the search space and the identification of strain communities of any potential size.

2nd set of experiments:

The diversity graphs were reconstructed for each carbon source. Mutants that cannot grow in a given environmental condition (lethal mutations) are excluded from the corresponding graph. The quantitative description of the metabolic difference with respect to by-production (see 4.2) produces a naturally weighted undirected graph with weight values in [0, 1], which is actually invariant to the initial amount of the carbon source. The binary representations of

the graphs are produced using a threshold of 0.6 (60% relative 'difference') in their corresponding weighted graphs. The specific threshold choice enforces the study of relations between mutants that are highly different.

Structural analysis is performed. Specifically, graph properties such as the centrality, the assortativity, the clustering coefficient and the maximum clique size (see 2.2) are evaluated for the variety of the growth conditions. Apart from the clique identification problem, which uses the binary and structurally compressed representation (see 4.2.4) of the diversity graph, the rest properties are applied on both the weighted and binary representations. Finding all cliques in a graph is computationally expensive in general. However, the graph allows significant compression to the number of nodes (structural compression) and the clique identification problem can be solved fast particularly when efficient, exact methods proposed elsewhere are used [92]. The properties of the nodes of the diversity graphs are also investigated with respect to the evolutionary trait of the corresponding deleted gene. The Evolutionary Retention Index (ERI) presented in the study of Gerdes et al [39] is used to represent how conserved a specific gene is across various bacteria. The ERI is determined for each *E. coli* ORFs by calculating the fraction of genomes in the group that have an ortholog of the given ORF with the number of representative organisms equal to 33. Thus, ERI takes values within [0, 1] where 0 corresponds to genes unique to *E. coli* and 1 for omnipresent genes.

The purpose of this second study is multifold. The graph representation was initially developed for dimensionality reduction purposes that allow the identification of strain communities, which have the potential to differently shape the environment resulting in interesting emergent phenotypes. The exploration of the properties of the diversity graphs aims to comparative analysis and the identification of consistent metabolic behaviors across the different growth conditions. The questions that are addressed involve whether there are mutants of unique, novel phenotypes and if novel phenotypes are observed whether they are environmental-specific or not or which gene deletions are responsible for their metabolic capabilities.

3rd set of experiments:

The third set of experiments focuses on a subset of carbons sources including *glycolate*, *acetate*, *glycine*, *glucose, pyruvate* and *melibiose* and simulates all the potential strain communities (cliques) found in each diversity graph using the proposed multi competitor growth model. The functional compression of the diversity graph (see 4.2.4) is used in the simulations of the cliques of size greater than 2. Unless differently specified, all competitors in the composition of the potential community appear with equal initial frequency in the population and the same initial amount of 10mmol/lt of the carbon source is provided to the system for growth. The potential communities of each selected growth condition are studied with respect to their growth benefit, the involved metabolic interactions and the predictability of their growth performance. The aim of this part, which comprises the functional analysis of the graphs, is the identification of efficient strain communities, the conditions that lead to efficient exploitation of the environment, the relationships between metabolic diversity and efficiency and between community size and efficiency.

In the following sections, mutants are named after the name of the gene that has been deleted for simplicity reasons.

5.2 Exhaustive Analysis of the WT-KO$^-$ pairs in different carbons

Table 5.1 summarizes the conditions in which the co-growth of a specific mutant with the wild type cell shows superior group performance. The relative yield ($Yield_{rel}$) is calculated with respect to the performance of the homogeneous population of either the wild type or the

participating mutant ($gene_{S\_KO}$) depending on which of the two participant's performance is maximum. The absolute yield ($Yield_{abs}$) is calculated with respect to the mutant ($gene_{\_KO}$) of the best homogeneous performance in the certain environment. The contents of Table 5.1 are sorted according to the value of the absolute yield. Only examples of positive relative yield that quantitatively describe group benefit are shown. Half of the cases presented in Table 5.1 exhibit positive absolute yields. For these conditions, certain mutants appear to efficiently contribute in synergy with the wild type cells. The performance of these mutants in other growth conditions is presented in Table 5.2. Some best synergistic mutants in specific carbons conditions are found to be lethal in other conditions.

The first column of Table 5.1 shows the carbon source as abbreviated in the supplementary files of [200]. The rest of the Table 5.1 is organized in four parts. The first part (WT) depicts the growth performance ($BM_{\_WT}$) of the homogeneous wild type population. The second part (Best KO^-) presents the performance (BM) of the best homogeneous mutant population ($gene_{\_KO}$). The yield of the best mutant ($Yield_{WT1}$) which describes the relative difference of the performances with respect to the wild type homogeneous population growth is also shown. For the conditions with a beneficial synergistic strain, we also always find a mutant with superior performance with respect to the wild type. The third part (Best synergistic KO^-) shows the performance of the homogeneous mutant ($gene_{S\_KO}$) population that when grown with the wild type in the given environment performs better than the homogeneous population of both the specific mutant and the wild type growth. In most of the cases the performance (BM) of this mutant in a homogeneous population is worse than the wild type as depicted by the corresponding yield ($Yield_{WT2}$). The last part shows the performance of the synergy. BM_{S_WT} is the final biomass of the wild type population while BM_{S_KO} is the final biomass that the mutant ($gene_{S\_KO}$) produces when they both coexist in the certain environment. The relative and absolute yield of the synergy, are depicted in the last two columns. Only for the growth on *L argenine* (*arg_L*) the best homogeneous mutant is also the best synergistic mutant. For growth on *glycine* (*gly*) and *glycolate* (*glyclt*) no mutant is found to perform better than the wild type in homogeneous populations.

Heterogeneous cell populations can exhibit superior growth by exploiting their metabolic by-products with mutual benefit. When for example a specific mutant population mutually grows with the wild type population on *glycolate* as primal carbon source, it becomes more efficient than when it grows alone. Dynamic profiles of the exchange substrates reveal that the mutant population becomes efficient because of the availability of *formic acid*, a byproduct which only the wild type population is capable of producing. This metabolic exchange leads the whole system to superior performance. In the following, we present a detailed analysis of the synergistic growths on *glycolate, citrate* and *pyruvate*.

TABLE 5.1
CONDITIONS OF SUPERIOR GROUP PERFORMANCE – GROWTH ON CARBON SOURCE

CARBON SOURCE	HOMOGENEOUS POPULATION							HETEROGENEOUS POPULATION			
	WT	Best KO⁻			Best synergistic KO⁻			Synergy			
	$BM_{\_WT}$	$gene_{\_KO}$	BM	$Yield_{WT1}$ (%)	$gene_{S\_KO}$	BM	$Yield_{WT2}$ (%)	BM_{S_WT}	BM_{S_KO}	$Yield_{rel}$ (%)	$Yield_{abs}$ (%)
glyclt	0.1025	-	-	-	'b2276'	0.0944	**-7.95**	0.085116	0.025721	**8.07**	**8.07**
gly	0.0660	-	-	-	'b2276'	0.0399	**-39.4**	0.06132	0.006615	**2.89**	**2.89**
cit	0.3585	'b0331' 'b0333' 'b0334'	0.3603	**0.49**	'b0728'	0.3331	**-7.08**	0.23057	0.1389	**3.05**	**2.54**
pyr	0.2365	'b3403'	0.2417	**2.19**	'b0721'	0.1313	**-44.8**	0.217158	0.028121	**3.71**	**1.48**
ser_D	0.2375	'b3403'	0.2433	**2.44**	'b0721'	0.1362	**-42.6**	0.218683	0.027194	**3.53**	**1.06**
arg_L	0.6223	'b1744'	0.6544	**5.16**	'b1744'	0.6544	**5.16**	0.36028	0.30053	**0.97**	**0.97**
4abut	0.5206	'b1849'	0.5211	**0.09**	'b0451'	0.3982	**-23.5**	0.39198	0.13054	**0.36**	**0.27**
melib	1.3837	'b1602'	1.4142	**2.20**	'b2276'	1.0878	**-21.4**	1.1073	0.28443	**0.58**	**-1.58**
tre	1.3883	'b1602'	1.4172	**2.08**	'b1241'	1.3788	**-0.68**	0.76453	0.62803	**0.31**	**-1.74**
sucr	1.3883	'b1602'	1.4172	**2.08**	'b1241'	1.3788	**-0.68**	0.76453	0.62803	**0.31**	**-1.74**
malt	1.3883	'b1602'	1.4172	**2.08**	'b1241'	1.3788	**-0.68**	0.76453	0.62803	**0.31**	**-1.74**
mnl	0.7782	'b1602'	0.7965	**2.34**	'b2276'	0.6283	**-19.2**	0.65929	0.12161	**0.34**	**-1.95**
akg	0.3933	'b4015'	0.4147	**5.44**	'b0721'	0.1343	**-65.8**	0.32267	0.078385	**1.97**	**-3.29**
glu_L	0.4764	'b4015'	0.5089	**6.83**	'b3236'	0.4774	**0.21**	0.26654	0.21778	**1.66**	**-4.83**

TABLE 5.2

	Growth performance of homogeneous mutant populations					
CARBON	**'b2276'**	**'b0728'**	**'b0721'**	**'b1744'**	**'b0451'**	**$BM_{\_WT}$**
glyclt	0.0945	0.0862	0.0861	0.0688	0.0030	0.1025
gly	0.0399	0.0660	0.0030	0.0660	0.0030	0.0660
cit	0.0116	0.3331	0.1314	0.3585	0.0030	0.3585
pyr	0.1910	0.2033	0.1313	0.2365	0.0030	0.2365
ser_D	0.1912	0.2040	0.1362	0.0268	0.1621	0.2375
arg_L	0.3231	0.6157	0.1385	0.6544	0.0128	0.6223
4abut	0.3131	0.5115	0.0030	0.5206	0.3982	0.5206

	Mutation has no effect on growth performance
	Mutant doesn't grow: Essential gene
	Best synergistic mutant

5.2.1 Case studies exhibiting superior growth

A. Aerobic growth on *glycolate*

Systematic examination of all possible single gene deletions reveals the mutant 'b2276' to have superior synergistic performance under growth on *glycolate*. The 'b2276' gene is related to *NADH dehydrogenase* involved in the respiratory chain of bacteria. The dynamic evolution of biomass for both the homogeneous and heterogeneous populations is shown in Fig. 5.1. The specific mutation affects the growth rate, thus a slower growth is observed in the homogeneous mutant population than that of the wild type under the same initial conditions. As a result, smaller amount of biomass is produced by the mutant population. However, the mutual growth of the wild type together with the mutant population shapes the environment in a beneficial way leading to about 8% higher performance (Table 5.1).

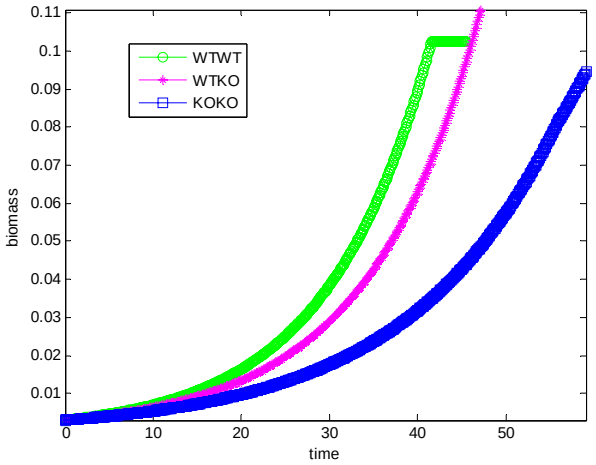


Figure 5.1 Predictions of biomass in batch cultures of homogeneous wild type (WTWT), homogeneous mutant (KOKO) and heterogeneous (WTKO) cell populations grown on glycolate. The mutant of best synergistic performance is produced by the knockout of gene 'b2276'. Time is in hours and the biomass is in gram [dry weight]/lt.

This superior performance of synergism is also reflected in the ratio of biomass rate to *glycolate* uptake rate (Fig. 5.2) that measures the efficiency of the cell system to convert the main carbon source available to essential biomass. Why the mutant population becomes so efficient when growing together with the wild type? Dynamic profiles of the exchange substrates reveal that the mutant population becomes efficient because of *formic acid*, a byproduct which only the wild type population is capable to produce. The increase and decrease of the biomass efficiency that is observed for the first 11.3 h (Fig. 5.2), in the mutant and the wild type population, respectively, is due to the *acetate* metabolism (production and consumption) that only the mutant population produces. *Acetate* is an essential common byproduct which can also be consumed by the cells. In other words, the beneficial performance of this particular pair is mainly due to the exchange of *acetate* and *formic acid* (Fig. 5.3).

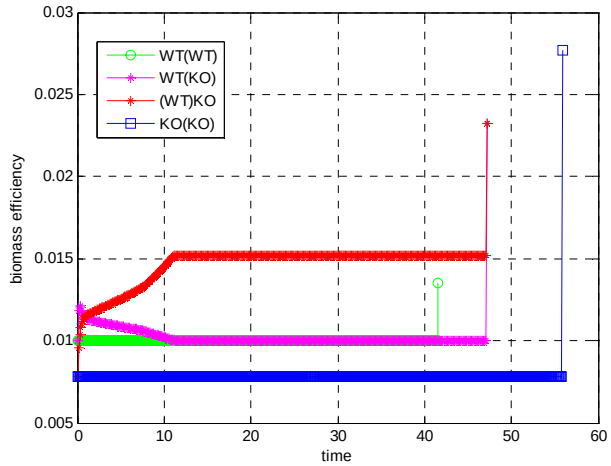


Figure 5.2 Predictions of biomass (growth) efficiency with respect to *glycolate* uptake in batch cultures of homogeneous wild type (WT-WT), homogeneous mutant (KO-KO) and heterogeneous cell populations grown on *glycolate* as the main carbon source. The figure shows how the WT population depicted as WT-KO and the KO population depicted as KO-WT performs in heterogeneous growth. The mutant of best synergistic performance is produced by the knockout of gene 'b2276'. Time is in hours and the biomass efficiency is in gram [dry weight]/mmol.

Apart from this initial *acetate* metabolism phase, the reasons that make synergistic growth on *glycine* also superior are similar. Furthermore, under homogeneous growth no mutant was found to be superior when compared to the wild type performance under growth on either *glycolate* or *glycine*.

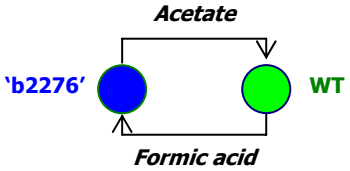


Figure 5.3 The cross-feeding interactions involved during the co-growth of WT with the mutant 'b2276' on limited *glycolate* lead the coexistence to supreme growth.

B. Aerobic growth on *pyruvate*

The metabolic pathways that are activated for the optimal consumption of *pyruvate* lead to the production of *acetate*. A mutant with the capacity of producing *acetate* at high concentrations is a potentially efficient partner to mutually grow with. However, *acetate* has a low growth rate when compared to the growth rate of *pyruvate*, thus a slight benefit might arise that depends on the growth efficiency of the mutant.

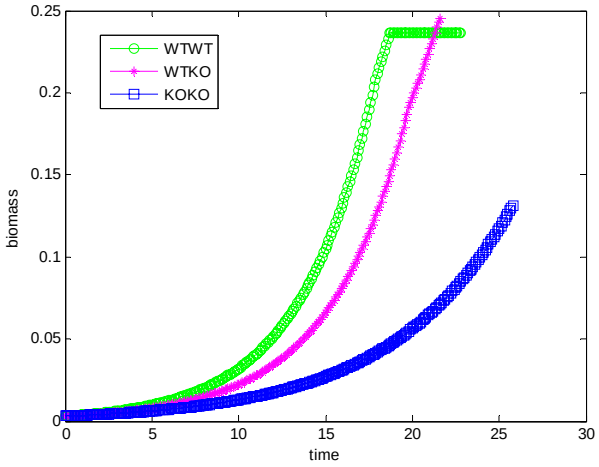


Figure 5.4 Predictions of biomass in batch cultures of homogeneous wild type (WTWT), homogeneous mutant (KOKO) and heterogeneous (WTKO) cell populations grown on *pyruvate*. The mutant of best synergistic performance is produced by the knockout of gene 'b0721'. Time is in hours and the biomass is in gram [dry weight]/lt.

Simulations on *pyruvate* show that the mutant generated after deleting the gene 'b0721' has superior synergistic performance as shown in Table 5.1. This mutant produces similar effects under other conditions such as growth on *D serine* (ser_D), *2 oxoglutarate* (akg) and *L glutamate* (glu_L). The gene 'b0721' codes for the enzyme *succinate dehydrogenase* which is involved in two metabolic reactions of different pathways the citrate cycle and the oxidative phosphorylation.

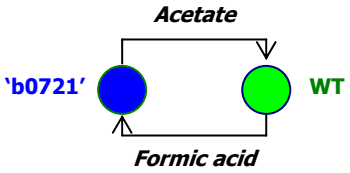


Figure 5.5 The cross-feeding interactions involved during the co-growth of WT with the mutant 'b0721' on limited *pyruvate*.

The growth profiles of the heterogeneous wild type - mutant populations as well as of the homogeneous wild type population show two exponential phases (Fig. 5.4). The first phase (rapid growth) coincides with the consumption of *pyruvate* while the second (slower growth) is mainly related to the consumption of the by-product *acetate* generated during the first phase. The mutant of 'b0721' gene produces *acetate* but it is not capable of consuming it. Thus, no exponential shift is observed in the growth profile of the homogeneous mutant population. Furthermore, the growth rate is significantly affected by the deletion of gene 'b0721' leading the homogeneous mutant population to poor performance as shown in Fig. 5.4. In synergy, the biomass ratio of the mutant to the wild type population at the end of growth is about 1/7.7. Surprisingly, even though the mutant contribution to the total biomass is minor in synergy and no synergistic benefit is observed during the first phase the acetate production efficiency of the mutant cells eventually results in a superior performance.

C. Aerobic growth on *L-arginine*

Dynamic simulations of aerobic growth on *L-arginine* reveal a mutant ('b1744') that at the end of homogeneous growth has produced more biomass than the wild type (~5% yield as shown in Table 5.1). Furthermore, the synergy between this mutant and the wild type populations is predicted to be even more efficient (~1% absolute and relative benefit) as

shown in Table 5.1. These observations give rise to two interesting questions. One concerns the reason that the homogeneous mutant population exhibits superior performance when compared to the homogeneous wild type growth. The second concerns the synergistic benefit.

The temporal concentration profiles reveal that the mutant of the knockout gene 'b1744' redirects the fluxes towards the production of *putrescine* with no significant effect on the growth rate. *Putrescine* is the intermediate product of the *arginine* metabolic pathway and comprises a compound exhibiting a wide range of applications in chemical industry [204]. Simulations show that the wild type population does not produce *putrescine*. Applying flux variability analysis (FVA) [190] as a method to identify reactions that are critical for the optimal fluxes on the initial conditions on the wild type cells gives zero flux variability related to the *putrescine* exchange reaction. This means that the metabolic path towards the production of *putrescine* is not useful with respect to the growth rate. However, in a dynamic environment in which the main substrate *L-arginine* will eventually get exhausted, the *putrescine* production plays an important role as it is used as secondary resource providing a long-term benefit to cell growth. These observations justify the superior performance of the mutant 'b1744' in a source limited environment.

Further experiments show that the presence of *putrescine* in the environment even at minor amounts (0.0002 mmol/l) redirects the pathways and constrains the uptake flux of *L-arginine* to 3.12 mmols/gram/h in contrast to 4.6 mmols/gram/h that is observed when no *putrescine* is present in the environment. In that way the synergistic environment is beneficial since the WT cell population efficiently consumes *L-arginine* with respect to the biomass it produces. A slight increase in growth rate (0.5% relative increase) is observed as well. *L-arginine* lasts longer allowing more biomass to be produced. When it is exhausted the system has already produced more biomass than the homogeneous mutant population. Growth ends when the remaining *acetate* and *putrescine* are consumed as well.

5.2.2 Frequency-dependent interactions

The way growth characteristics such as the maximum growth rates, the group performance and the relative fitness (as defined in section 1.3.4) of the coexisting genotypes depend on their initial frequencies is explored here. Two examples are shown; the first involves the co-growth of the best synergistic pair on *glycolate* whereas the second example corresponds to growth on *L-arginine*.

The co-growth experiments of the WT and the mutant 'b2276' cells evaluated for different initial ratios on limited *glycolate*, show that the group-benefit increases when the WT population is present at a lower initial ratio in the population (Fig. 5.6-Left). Maximum group performance is observed when the initial ratio for WT:'b2276' is 0.05:0.95. The fitness of the WT relative to the mutant b2276 is always greater than 1 for all different initial frequencies, however it is observed to increase when the WT is rare (Fig. 5.6-Right).

The maximum growth rate of the WT cells is higher relative to the mutant 'b2276' cells. The maximum growth rate is an important component of fitness in repeated competition experiments and evolution experiments [21]. Here, it is shown that the maximum growth rate of the WT is frequency-dependent and specifically it increases as its population becomes rarer (Fig. 5.7). Furthermore, the growth rate time profiles show that when the WT is rare, it achieves its maximum growth rate faster and this phase of efficient growth lasts longer relative to when it is in abundance. *Acetate* constrains the fluxes and is responsible for the period of time the WT population operates at maximum growth rate.

Simulations performed on limited *L-arginine* for different initial ratios of the WT and the mutant 'b1744' didn't show any considerable frequency-dependence of the growth characteristics (Fig. 5.8).

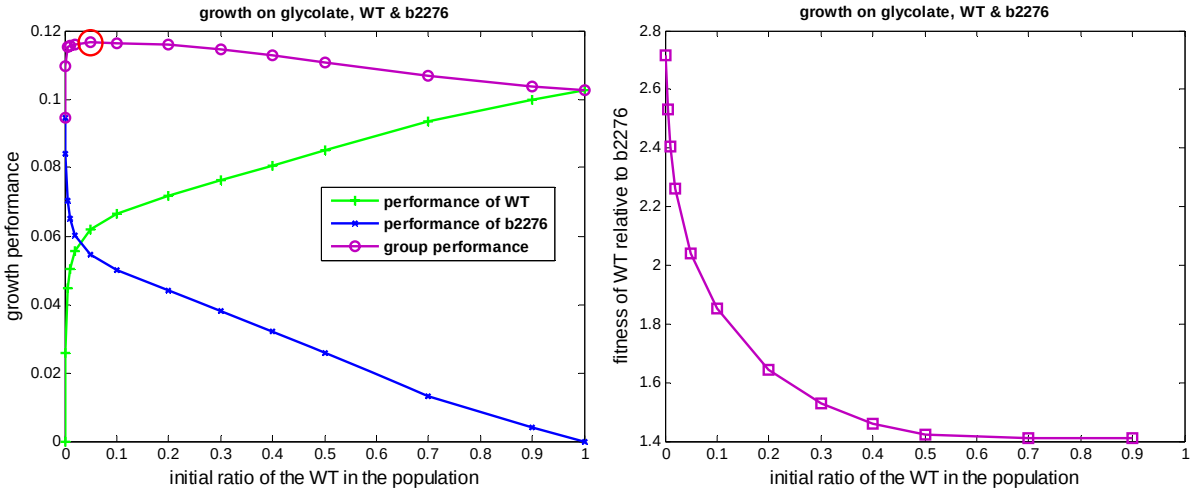


Figure 5.6 Co-growth of the WT and the mutant 'b2276' populations on limited *glycolate*. (Left) The growth performance of the WT, the mutant 'b2276' and the group is shown as a function of the initial frequency of the WT. The group performance is maximized when the initial population ratio is 0.05:0.95 for WT: 'b2276'. (Right) The fitness of the WT relative to the mutant b2276 is shown for different initial population ratios.

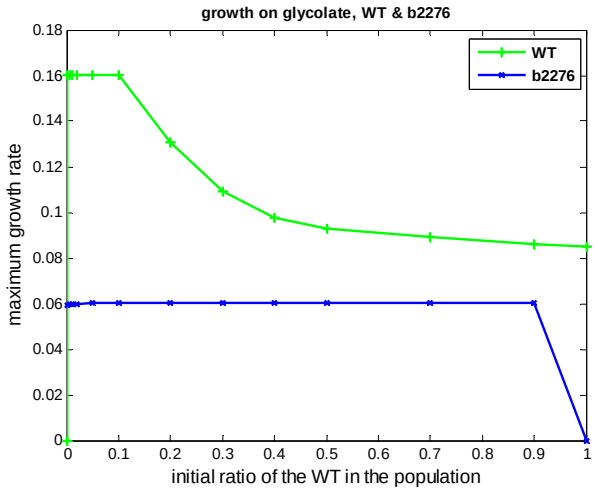


Figure 5.7 Co-growth of the WT and the mutant 'b2276' populations on limited *glycolate*. The maximum growth rate of the WT cells is frequency-dependent.

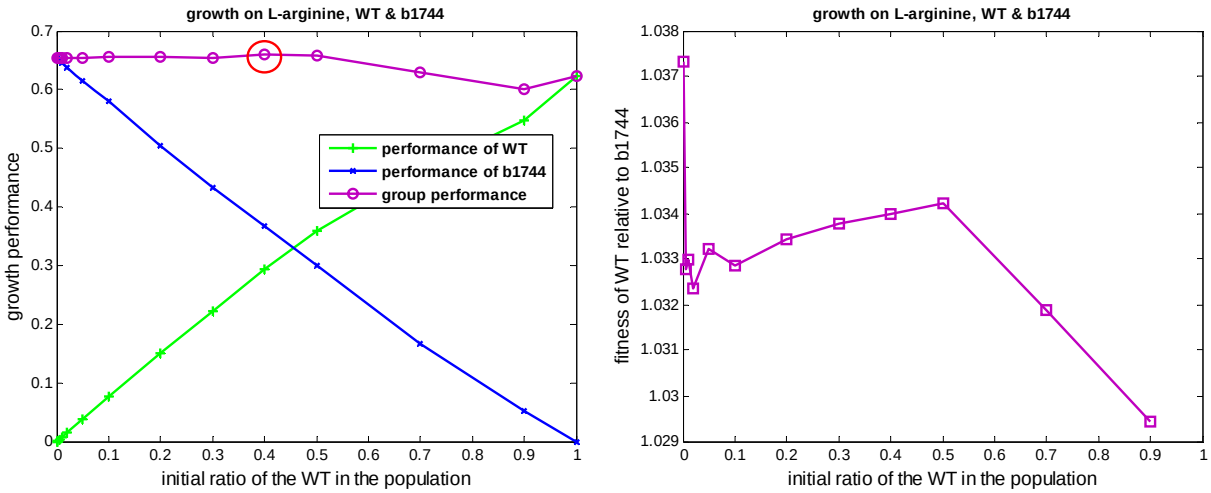


Figure 5.8 Co-growth of the WT and the mutant 'b1744' populations on limited *L-arginine*. (Left) The growth performance of the WT, the mutant 'b1744' and the group is shown as a function of the initial frequency of the WT. The group performance is maximized when the initial population ratio is 0.4:0.6 for WT: 'b1744'. (Right) The fitness of the WT relative to the mutant 'b1744' is shown for different initial population ratios.

5.2.3 Ecological-dependent interactions

The question that is addressed here is whether the group benefit, which involves cross-feeding interactions such as those observed when the WT and the mutant 'b2276' co-grow on limited *glycolate* depend on the concentration of the primal resource. This is important because it can determine the ecological opportunity under which the exchange of intermediate products of metabolism is essential for growth. Experiments reviewed in [10] have shown that ecological opportunity is critical for the evolution of polymorphism in bacterial communities.

The co-growth of the wild type and the mutant 'b2276' is examined for seven different initial concentrations of *glycolate.* Specifically, the concentration of 10 mmol/l are multiplied by the factor {0.1, 0.3, 0.5, 1, 2 10, 20}. Oxygen and ammonia, as well as other important inorganic metabolites that are initially provided in the growth medium are multiplied accordingly as well. In the previous section, it was observed that the group performance is frequency-dependent and that it is maximized when the initial ratio for WT:'b2276' is 0.05:0.95. At this initial ratio the cross-feeding interactions seem to play the most important role. The *acetate* that the mutant cells provide to the wild type cells enhances the growth performance of the wild type and is mainly responsible for the observed group benefit. Therefore, the variable concentration experiments are performed for initial ratios equal to 1:1 and 0.05:0.95 for WT:'b2276'.

The dependence of the relative group benefit on the initial concentration of *glycolate*, which is defined with respect to the growth performance of the homogeneous wild type population under the same initial conditions, is shown in Fig. 5.9. It is observed that the group-benefit decreases as the initial concentration of *glycolate* increases by considerable amounts. This suggests that cross-feeding interactions are more important when cells experience the resource limitation than when the food is in abundance in their environment. Furthermore, it is observed that the group-benefit is considerably decreased at low *glycolate* concentrations (less than 10mmol/l) when the initial population ratio is 0.05:0.95 for WT:'b2276'. At these low *glycolate* concentrations, the *acetate* provided by the mutant is not suffiecient for the wild type to exhibit high performance. Overall these findings suggest that cross-feeding plays an important role in promoting group benefit under certain ecological conditions that can support it.

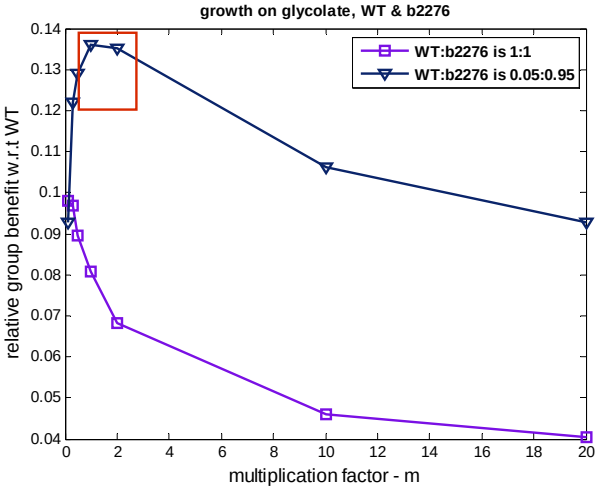


Figure 5.9 Ecological opportunity (red window) for the emergence of cross-feeding interactions and group benefit when WT and mutant-'b2276' cells co-grow on limited *glycolate*. The relative group benefit is shown as a function of the *glycolate* initial concentrations. The purple line corresponds to initial ratio 1:1 for WT:'b2276' whereas the dark blue line corresponds to initial ratio 0.05:0.95.

5.2.4 Flux variability

The optimum flux distribution (metabolic state) predicted by the FBA for a given set of constraints can either be unique (single solution) or an infinite number of equivalent solutions can exist (multiple solutions). Although the optimal solution space can reflect missing knowledge about the biological system, there is evidence in support of an optimal solution space that displays biologically meaningful information. This problem was analyzed in section 3.4.4.

In dynamic simulations, the optimal solution (single or multiple) varies depending on the constraints of the optimization problem change. If multiple solutions exist, each solution may differently shape the environment determining in that way its own (instant) optimum pathway over time. In this respect, the temporal predictions provided by the dynamic FBA describe one of these possible pathways the cell can follow as grows in a given dynamic environment. Exploring the dynamic solution space is beyond the scope of this study however, it is mentioned here because of its importance.

To understand how alternative optimal solutions can differently shape the growth medium, FVA is performed on the exchange reactions while varying the constraints of the problem. In selected examples including homogenous and heterogeneous population growth, FVA shows that during the metabolism of the main source the fluxes of the exchange reactions exhibit flux variability, which is of the order of 10^{-3}, whereas after the depletion of the main source several exchange reactions are observed to exhibit high flux variability of the order of 1.

5.2.5 Metabolic variability

All the growth efficient wild type-mutant pairs shown in Table 5.1 involved cross-feeding interactions. The exchange between the populations involved either a novel metabolite or an essential metabolite, which one of the populations was capable of providing it at high quantity. This suggests that the best synergistic mutant exhibits metabolic capabilities different than the wild type cell in the same growth conditions. This observation significantly reduces the search space of strain communities with the

TABLE 5.3

Carbon Source	Best WT-KO pair	Edge weight
glyclt	WT-'b2276'	1
gly	WT-'b2276'	1
cit	WT-'b0728'	1
pyr	WT-'b0721'	1
ser_D	WT-'b0721'	1
arg_L	WT-'b1744'	1
4abut	WT-'b0451'	1
melib	WT-'b2276'	0.665
tre	WT-'b1241'	1
sucr	WT-'b1241'	1
malt	WT-'b1241'	1
mnl	WT-'b2276'	0.904
akg	WT-'b0721'	1
glu_L	WT-'b3236'	0.995

potential to develop cross-feeding interactions, as shown in the next set of experiments. The maximal metabolic variability (edge weight) between the WT and the mutant determined from their homogeneous growth experiments as described in 4.2 is shown in Table 5.3. A value equal to 1 implies the presence of a novel metabolite.

5.2.6 Conclusions

In the first set of experiments this work investigates in a systematic, computational way all the wild type- mutant coexistences in a common growth medium under various single-carbon sources and compares the group performance with the performance of the corresponding homogeneous populations in the same initial conditions. The involved populations are allowed to exchange by-products through the environment they grow in, in order to exploit the common environment for their own benefit.

In the metabolic simulations, bacterial cells are modeled to operate in a greedy way that maximizes their growth rate at each time step. Unless the deleted gene corresponds to an inactive or redundant reaction, single gene knockout mutants have a more restricted metabolic network to operate than the wild type cells since the fluxes of the pathways that are associated with the deleted gene are constrained to zero. Thus, mutants have in general lower (or at most equal) growth rate than the wild type during the metabolism of the main carbon source. Nevertheless, two scenarios surprisingly appear.

First, it is possible to have homogeneous mutant population that exhibits superior growth performance when compared to the homogeneous wild type growth. Alternative pathways that produce byproducts essential for growth when the main substrate is exhausted are revealed. These strategies might sacrifice the instant maximal growth rate but eventually are proved to be more efficient with respect to the endpoint biomass concentration. Growth on secondary metabolites is in general slower than growth on the primal source due to thermodynamic constraints. For efficient strategies to emerge the benefit from the secondary resource metabolism must exceed the loss of growing at lower growth rate during the metabolism of the primal resource (positive balance). Such an example is the growth on *L-arginine*. The metabolic pathways toward the production of the intermediate product *putrescine* are not optimal with respect to the growth rate under the initial conditions of a medium rich in *L-arginine* (Flux Variability Analysis experiments). However, these pathways show endpoint benefits. When the main substrate is exhausted, *putrescine* plays an essential role leading the system to superior performance. Growth simulations on *pyruvate* also revealed the mutant coming from the deletion of the gene 'b3403' as the most efficient of all mutants and the wild type. The deleted gene encodes the enzyme *phosphoenolpyruvate carboxykinase* and is involved in the *anaplerotic reaction* subsystem. The specific mutant exhibits maximum growth rate over time that is lower than the maximum growth rate possible; however since it is capable of producing a large concentration of *acetate* eventually more biomass is produced.

Second, group benefit may arise even in cases where the individual growth performance of the mutant cells is smaller in coexistence where there is resource competition between the strains than when the mutant cells grow alone (non competitive environment). Under the previous consideration, heterogeneous cell populations can exhibit superior growth by exploiting their metabolic by-products with mutual benefit. In several growth conditions the heterogeneous growth performance is superior to the homogeneous growth performance of any mutant (positive absolute benefit) which implies that for the specific conditions more competitors can better utilize the resources than a monomorphic population.

All the growth efficient wild type-mutant pairs that are identified in this set of experiments involved cross-feeding. It was also shown that the best synergistic mutant exhibits metabolic capabilities with respect to by-production which were different than the metabolic capabilities of the wild type cell in the same growth conditions. These results are in consistence with

observations occurring in nature. Cross-feeding interactions have been observed to emerge between evolved strains in several long-term evolution experiments on a single limited resource with *E. coli* while the evolved strains significantly differed with respect to their gene expression patterns and their metabolic capabilities (reviewed in chapter 1).

It was also shown that the initial population ratio of the competitors can play an important role in the outcome of the competition by shaping their growth characteristics, which include the maximum growth rate, the group performance and the relative fitness. Furthermore, it was shown that a high enough amount of the primal resource (e.g. *glycolate*) in the medium can decrease the group benefit whereas a low enough amount can decrease the cross-feeding effect determining in that way an ecological window in which cross-feeding can promote group benefit more effectively.

5.3 Metabolic Diversity Graphs

The diversity graphs were reconstructed for each growth condition based on the simulated metabolic capabilities of each mutant. Graph properties such as the centrality, the assortativity, the clustering coefficient and the maximum clique size are presented. Consistent patterns of metabolic behavior are explored here as well.

5.3.1 Examples of reconstructed networks

The binary representations of few examples after structural compression that maps all the structurally identical nodes onto a super node are visualized in the following. The highly connected nodes (depicted in red) correspond to mutants of unique phenotype with respect to by-production when compared with the other mutants of the graph. On the other hand, the super node (depicted in black) included mutants of similar metabolic capabilities. The wild type is also a member of the super node.

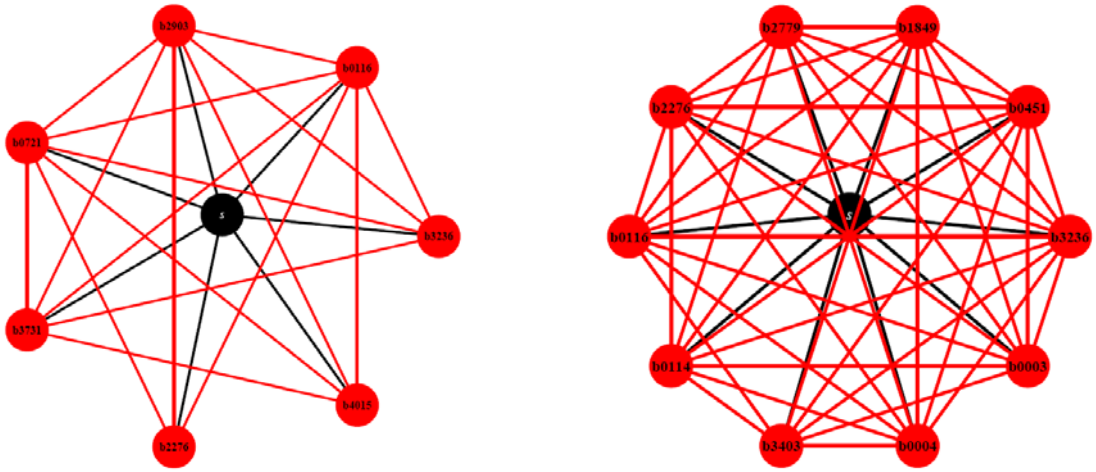


Figure 5.10 (Left) The binary representation of the diversity graph for growth on *pyruvate*. Seven highly connected nodes are revealed. 373 nodes comprise the super-node. (Right) The binary diversity graph of *4-aminobutanoate*. 10 nodes are highly connected whereas the super-node consists of 372 nodes.

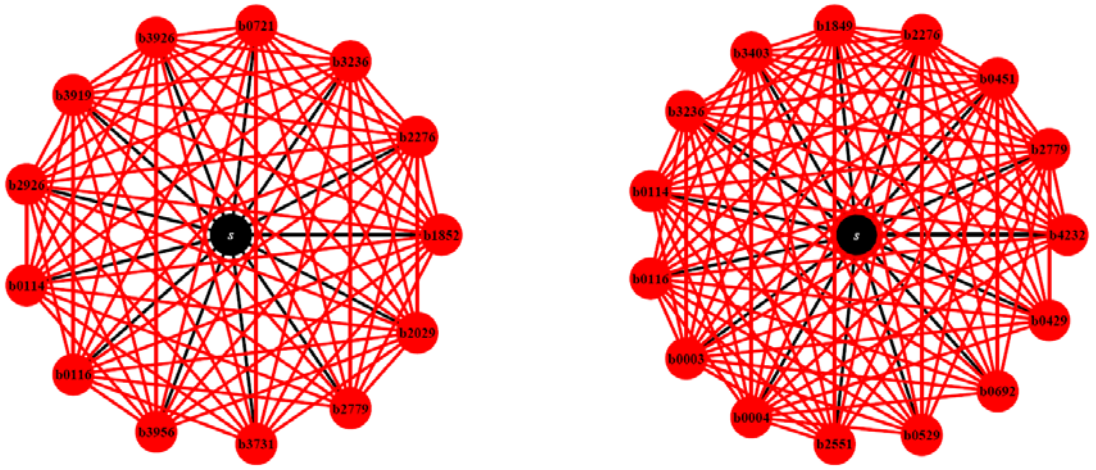


Figure 5.11 (Left) The binary diversity graph of *D glucose*. The graph consists of 13 highly connected nodes and the super-node includes 373 nodes. (Right) The binary diversity graph of *putrescine* consists of 380 nodes of which 15 are highly connected.

Not all diversity graphs can be visualized like that. However, as it is analyzed in the next paragraphs most graphs are characterized by a relatively small set of central nodes and a large set of nodes, which are disconnected with each other and connected with the central nodes.

5.3.2 Structural Analysis

5.3.2.1 Edge-weight distribution

The edge weights of the diversity graph encode the maximum difference that is observed between two genetically different mutants over all their products of metabolism with respect to their by-production efficient. Edge weight equal to 1 indicates that at least one mutant produces a metabolite that is novel for the other mutant. Edge weight equal to 0 indicates that both mutants produce the same metabolites at the exact same amount.

The weight distributions of the diversity graphs (Fig. 5.12) show that most edges (above the 70% for most conditions) are attributed with low weight values ($w < 0.1$), which indicates that most mutants have similar metabolic capabilities with each other with respect to by-production. However, few edges (below the 20%) are observed to take particularly high weight values ($w > 0.9$), which correspond to mutants of highly different metabolic capabilities with each other. Edges of weight values in-between ($0.1 < w < 0.9$) are found to be less than 10% in the graph. Exception to the generally observed skewed weight distribution comprises the diversity graph that corresponds to growth on *adenosine*, which exhibits an almost uniform weight distribution implying high metabolic variability in response to genetic perturbations.

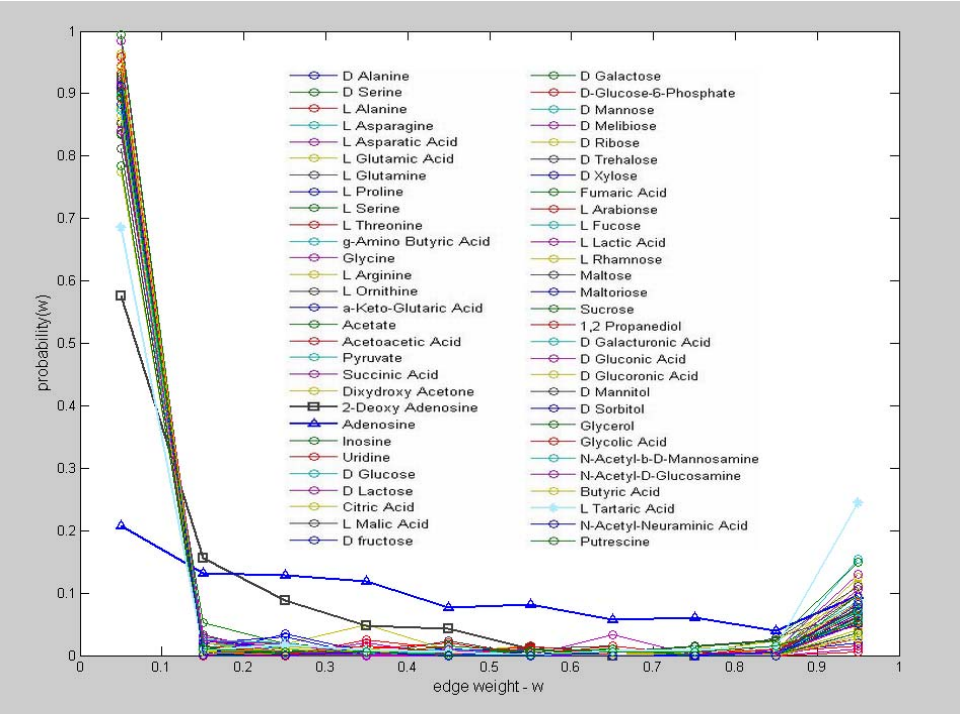


Figure 5.12 Edge weight distribution of the (unconstrained) diversity graph of each different carbon-source condition. Most edges have low weight values for most carbon conditions. Exceptions are growth on *Adenosine* and *2-Deoxy Adenosine*.

5.3.2.2 Strength centrality distribution

The strength centrality captures the importance of a node in the graph by taking into account the strength (weight values) of its connections (see 2.2.7.2). The normalized strength centrality takes values within $[0,1]$, where 0 implies a disconnected node and 1 a node connected with all the rest nodes of the graph with weights equal to 1.

In the diversity graphs, the highly central nodes indicate mutants of considerably different metabolic capabilities with respect to by-production from the rest mutants of the graph and highly non-central nodes correspond to mutants which mostly exhibit similar metabolic capabilities with the mutants they are connected with.

The strength-centrality distributions (Fig. 5.13) show a high percentage of nodes (above 80%) being non-central whereas few nodes (below 10%) appear to be highly central having strength centrality greater than 0.9. The skewed centrality distributions of the diversity graphs imply that when perturbations in the enzyme-coding genes take place in a cell, the metabolically similar responses or metabolic redundancies with respect to by-production are by far more common cellular behaviors than the different metabolic behaviors for most of the carbon conditions we have examined. The diversity graph of *adenosine* is observed to comprise an exception exhibiting a broader strength centrality distribution than the rest carbon-source graphs. The information that is encoded in the strength centrality of the nodes of the graphs across the various growth conditions and with respect to the evolutionary trait of the involved genes is further explored in section 5.3.3.

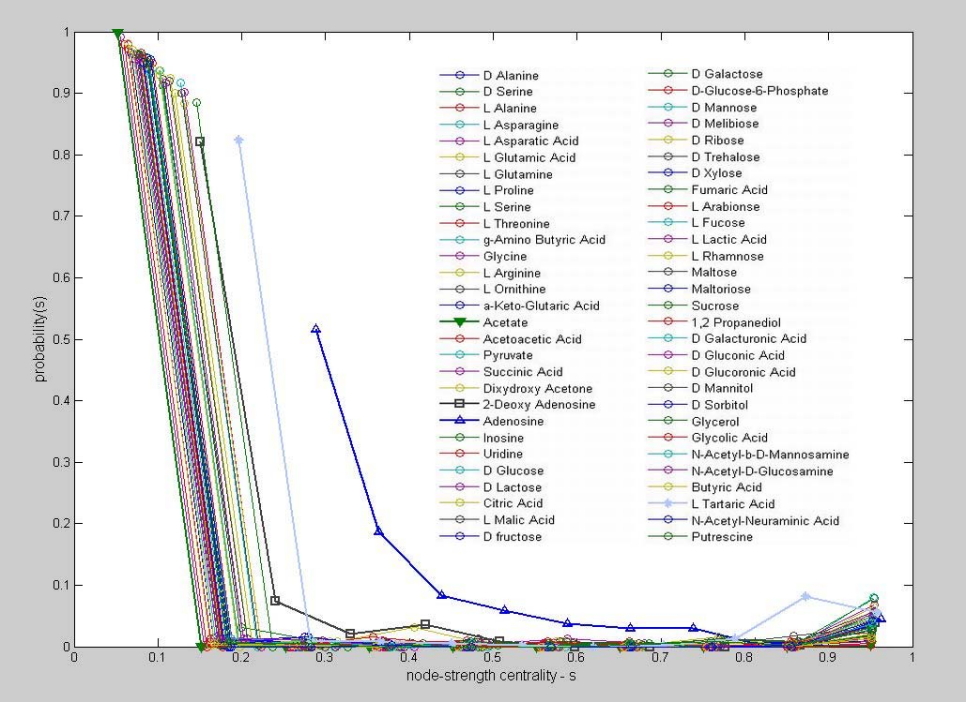


Figure 5.13 Node strength centrality distribution of the (unconstrained) diversity graph of each different carbon-source conditions. Most nodes have low centrality, whereas few nodes are observed to have considerably high strength centrality. The diversity graph of *adenosine* exhibits a broader strength centrality distribution than the rest graphs.

The number of the highly central mutants affects the centrality value of the non-central mutants since these two groups of mutant phenotypes are interdependent. The non-central mutants are actually strongly connected with the few central mutants that exist in the graph and are loosely connected with each other because they comprise metabolically redundant mutants of similar metabolic capabilities with each other.

It is important to mention here that if the weights are drawn randomly from the standard uniform distribution then the weight distribution would be uniform whereas the strength distribution would follow a Gaussian distribution of mean equal to 0.5 in the limit of large number of nodes (central limit theorem), which means that a node has equal probability to be connected or disconnected with another node.

5.3.2.3 Strength centrality – Clustering coefficient

If the node's strength centrality is plotted as a function of the node's clustering coefficient, it can provide insight into which nodes are actually responsible for the formation of clusters in the graph and the generation of potential bacterial communities.

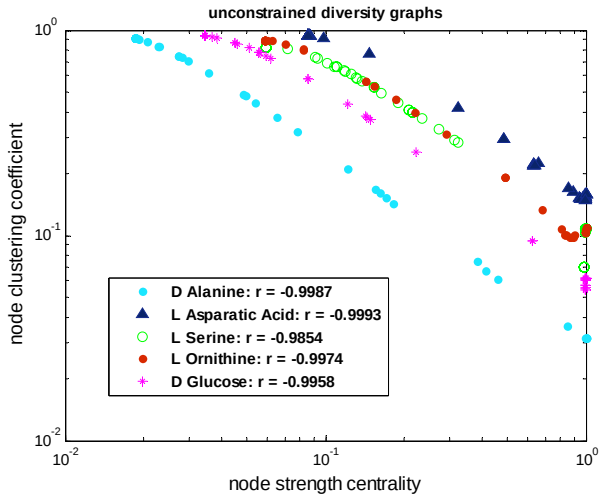


Figure 5.14 The node's clustering coefficient with respect to the strength centrality for five different growth conditions approximates a straight line on a log-log plot. The correlation coefficient values – r are shown (p-value = 0 in all cases). Highly central nodes have low clustering coefficient whereas the nodes of significantly low strength centrality appear to have high correlation coefficient value.

The plots that are shown in Fig. 5.14 correspond to a few selected growth conditions although almost all carbon conditions produce similar plots. It is observed that the clustering coefficient exhibits a non-linear dependence on the strength centrality. The relation between clustering coefficient and strength centrality approximates a straight line on a log-log plot. Highly non-central nodes (low strength centrality) exhibit clustering coefficient, which is considerably high (close to one). On the other hand, as the strength centrality of a node increases its clustering coefficient decreases so that the highly central nodes exhibit clustering coefficient close to zero.

The non-linear dependence that is observed between the two metrics is a direct consequence of the particular structure that characterizes the diversity graphs, which is close to a star topology with relatively few central nodes (hubs) forming the only highly clustered area in the graph. The central nodes are connected with a loosely connected area of relatively high size, which corresponds to the redundant group of mutants. Thus, the redundant group is part of the highly clustered areas and therefore their clustering coefficient is high. The mean clustering coefficient of the network remains however at a high level (Table 5.7) due to the fact that most of the nodes of the graph have low strength centrality.

In the following the graph of *D alanine* is used as an example to show that as the number of the central nodes decreases, the exponential decay is faster and eventually drops to (0, 0) corresponding to a disconnected graph (Fig. 5.15). On the other hand, as the highly clustered area of the diversity graph increases in size, the exponential decay is slower, the dependence becomes linear and eventually reaches the (1, 1) which corresponds to a complete graph (clique).

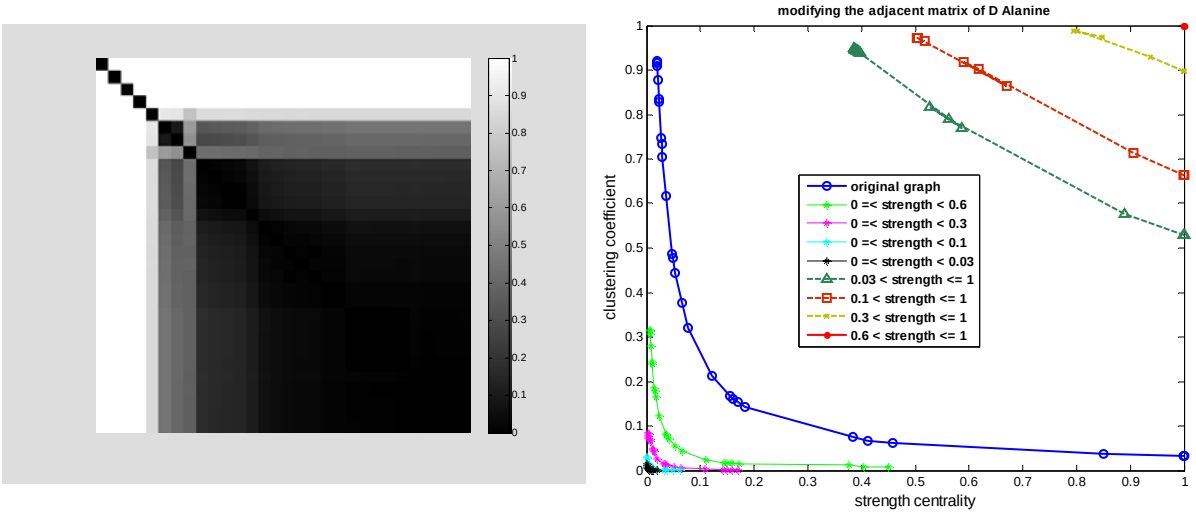


Figure 5.15 (Left) The adjacency matrix of *D Alanine* sorted by decreasing strength centrality. The first 30 nodes of the 383 nodes of the graph are presented here. The grey scale corresponds to the weight values of the edges of the graph. (Right) The clustering coefficient as a function of the strength centrality is shown, when different set of nodes based on their (original) strength centrality are selected. When the number of the non-central nodes decreases, the dependence becomes linear.

5.3.2.4 Shortest Paths

To express distance between the edges of the diversity graphs, a monotonic non-increasing function is applied as described in section 2.2.3. Edges with weight value equal to zero have infinite distance, whereas edge weights equal to 1 have distances equal to 1, expressing direct links.

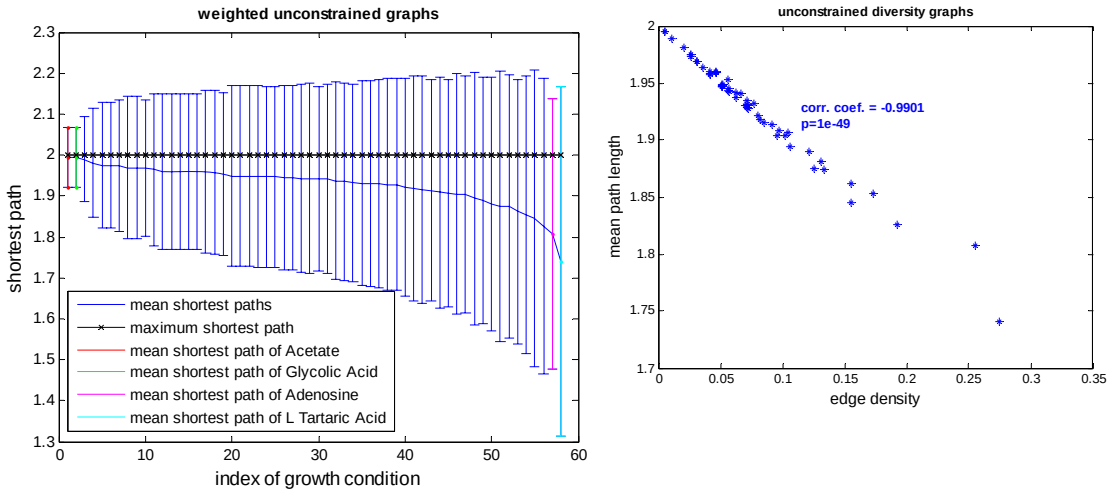


Figure 5.16 (Left) Shortest path lengths found in each diversity graph. The graph diameter equals to 2 in all cases. The mean shortest path length varies between 1 (direct link) and 2 (indirect link) whereas it is observed to be closer to 2 reflecting the indirect connectivity of the redundant group. (Right) The higher the edge density the closer to 1 is the mean path length.

It is observed that the longest shortest path (graph diameter) of all diversity graphs is equal to 2, which means that a node is reachable to another in at most two links. Therefore, the shortest path lengths in a diversity graph actually vary between 1 and 2. The mean path length represents the average over the shortest paths between all pairs of nodes and it is observed to be closer to 2 in all growth conditions. This indicates that the number of the

indirect paths is more than the direct, which is in correspondence with the low edge densities observed in the binary representation of the diversity graphs. A significant ($p = e^{-49}$) negative correlation is observed between mean path length and edge density. The nodes that correspond to the redundant group of mutants of high metabolic similarity and consequently high distance are connected with each other indirectly through the highly central nodes of the graph.

5.3.2.5 Network centrality

Most growth conditions independently on their different edge definitions and representations (see 4.2.2 and 4.2.3) generate diversity graphs that are highly centralized. As presented in Table 5.4, the mean value of the network centrality over all growth conditions is close to 1, which indicates that the diversity graphs are dominated by few highly central nodes (hubs) and thus exhibit a more star-like topology. This observation further implies that the metabolic network exhibits high metabolic robustness to genetic perturbations with respect to by-production under most of the growth conditions tested.

Table 5.4: Network Centrality over different carbon sources

WEIGHTED GRAPHS	mean Centrality	std Centrality	min-max Centrality	Carbon of min Centrality	Carbon of max Centrality
constrained	0.9266	0.0530	0.6221 – 1	'adenosine'	'acetate', 'glycolate'
unconstrained	0.9165	0.0637	0.6147 – 1	'adenosine'	'acetate'
s-unconstrained	0.8618	0.0799	0.4870 – 1	'adenosine'	'acetate'

BINARY GRAPHS	mean Centrality	std Centrality	min-max Centrality	Carbon of min Centrality	Carbon of max Centrality
constrained	0.9386	0.0396	0.7535 – 1	'adenosine'	'acetate', 'glycolate'
unconstrained	0.9289	0.0543	0.7297 – 1	'L Tartaric Acid'	'acetate', 'glycolate'
s-unconstrained	0.8694	0.0726	0.5865 – 1	'adenosine'	'acetate', 'glycolate'

Acetate generates diversity graph of centrality equal to 1 in all its different edge definitions and representations. *Glycolate* also exhibits network centrality equal to 1 in its binary graph representation. One central node is actually present in these graphs, which corresponds to a mutant that is metabolically different from all the rest mutants of the graph. On the other hand, growth on *adenosine* exhibits the lowest centrality of value close to 0.6 implying a system of lower metabolic redundancy. In general, the network centralities of the constrained and the unconstrained graphs behave similarly for most growth conditions. The super-unconstrained graphs are slightly less central than the constrained and the unconstrained graph cases because of the addition of a considerable number of edges. The binary representations of the diversity graphs exhibit in general behavior similar with the behavior of the weighted representations with respect to centrality implying that the selected threshold value preserves the information encoded in the graph.

5.3.2.6 Network assortativity coefficient

When the centrality of each node in the network is taken as a node property, the assortativity coefficient is observed to take a negative value close to -1 for the most of the growth conditions showing that the diversity graphs are highly disassortative. The observed disassortative mixing is a direct consequence of the structure of the diversity graphs where (few) highly central nodes are connected with the (many) non-central nodes. Table 5.5 shows the mean, the standard deviation as well as the minimum and maximum assortativity values over all graphs. Among all growth conditions, less disassortative is the diversity graph, which corresponds to the carbon source *adenosine*. The binary diversity graphs of *acetate* and *glycolate* have assortativity coefficient equal to -1, which is a direct consequence of their

star topology. Since the binary representation strengthens the divergence between the central and the non-central nodes the assortativity coefficient is observed to be affected towards higher assortativity values.

Table 5.5: Assortativity Coefficient (based on strength) over different carbon sources

WEIGHTED GRAPHS	mean Assortativity	std Assortativity	min-max Assortativity	Carbon of min Assortativity	Carbon of max Assortativity
constrained	-0.8006	0.120	-1 – -0.3112	'acetate', 'glycolate'	'adenosine'
unconstrained	-0.7868	0.110	-1 – -0.3041	'acetate'	'adenosine'
s-unconstrained	-0.8645	0.103	-1 – -0.3774	'acetate'	'adenosine'

BINARY GRAPHS	mean Assortativity	std Assortativity	min-max Assortativity	Carbon of min Assortativity	Carbon of max Assortativity
constrained	-0.9686	0.0359	-1 – -0.7257	'acetate', 'glycolate'	'adenosine'
unconstrained	-0.9638	0.0399	-1 – -0.7212	'acetate', 'glycolate'	'adenosine'
s-unconstrained	-0.9392	0.0429	-1 – -0.6930	'acetate', 'glycolate'	'adenosine'

The assortativity coefficient is also investigated with respect to the Evolutionary Retention Index (ERI) of the deleted gene that characterizes the mutant-node. The higher the conservation value of the gene under deletion is across divergent bacteria, the lower becomes the probability of the specific mutant to evolve. A disassortative mixing is observed for most carbon conditions as shown in Table 5.6 indicating a tendency of potential cross-feeding interactions to be developed between mutants derived from the deletion of a highly conserved gene and mutants derived from the deletion of a less conserved gene. This observation is important since it suggests that when the cell loses an essential, non-lethal, highly conserved gene then it can metabolically interact with another mutant who is not unlikely to evolve in a population and get rescued from extinction.

Table 5.6: Assortativity Coefficient (based on ERI) over different carbon sources

WEIGHTED GRAPHS	mean Assortativity	std Assortativity	min-max Assortativity	Carbon of min Assortativity	Carbon of max Assortativity
constrained	-0.2318	0.1106	-0.4432 - -0.0070	'acetoacetic acid'	'adenosine'
unconstrained	-0.2251	0.1117	-0.4431 - -0.0064	'acetoacetic acid'	'adenosine'
s-unconstrained	-0.1661	0.0774	-0.2913 - -0.0096	'acetoacetic acid'	'adenosine'

BINARY GRAPHS	mean Assortativity	std Assortativity	min-max Assortativity	Carbon of min Assortativity	Carbon of max Assortativity
constrained	-0.2439	0.1194	-0.5082 - -0.0098	' dixydroxy acetone '	'L arginine'
unconstrained	-0.2365	0.1219	-0.5014 - -0.0097	'dixydroxy acetone'	'adenosine'
s-unconstrained	-0.1701	0.0789	-0.3030 - -0.0137	'acetoacetic acid'	'adenosine'

5.3.2.7 Network clustering coefficient

The mean clustering coefficient values of the diversity graphs indicate an overall high tendency of the nodes to form or participate in clusters. The diversity graph of *glycine* shows the highest value of the clustering coefficient (Table 5.7). The diversity graphs of the *acetate* and *glycolate* on the other hand, have clustering coefficient equal to 0 or very close to 0, in all their diversity graph definitions, which is consistent with their start-like topology that does not allow other connections but those with the single central node. The diversity graphs, as mentioned previously, are highly centralized, consisting of few highly central nodes that are actually connected to many highly non-central nodes. Given this structural information, the high values of the network clustering coefficient imply that the central nodes are highly connected with each other forming a highly clustered area in the graph. The exact

dependence between clustering coefficient and centrality is shown previously in section 5.3.2.3.

Table 5.7: Network Clustering Coefficient over different carbon sources

WEIGHTED GRAPHS	mean Clustering	std Clustering	min-max Clustering	Carbon of min Clustering	Carbon of max Clustering
constrained	0.7966	0.1776	0 – 0.9946	'acetate', 'glycolate'	'glycine'
unconstrained	0.8207	0.1449	0 – 0.9695	'acetate'	'glycine'
s-unconstrained	0.7611	0.1688	0 – 0.9695	'acetate'	'glycine'

BINARY GRAPHS	mean Clustering	std Clustering	min-max Clustering	Carbon of min Clustering	Carbon of max Clustering
constrained	0.7932	0.2039	0 – 0.9946	'acetate', 'glycolate'	'glycine'
unconstrained	0.8126	0.2023	0 – 0.9946	'acetate', 'glycolate'	'glycine'
s-unconstrained	0.7415	0.2024	0 – 0.9946	'acetate', 'glycolate'	'glycine'

5.3.2.8 Maximum cliques

The number of the highly central nodes provides an indication of the different metabolic capabilities of the system in a given growth condition. However, the connectivity of the highly connected nodes with each other is also important. The maximum clique size that can be found in a diversity graph reflects the actual number of the different metabolic patterns that can emerge in the system under the given growth condition and the specified, allowable genetic perturbations. Based on the assumption that cross-feeding requires metabolic variability in the population in order to take place, cliques correspond to strain communities of the potential to develop cross-feeding interactions and the maximum clique size defines the upper bound of the strain communities.

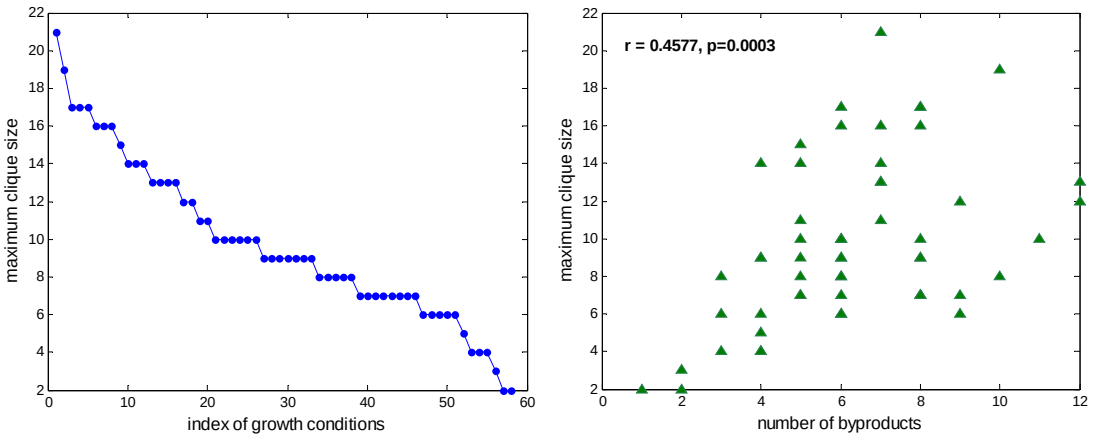


Figure 5.17 (Left) The maximum clique size as found in the binary representation of the diversity graph of each growth condition. (Right) The number of the metabolites that are by-produced in each growth condition with respect to the maximum clique size.

The maximum cliques that are presented here correspond to the binary representation of the unconstrained diversity graphs. The maximum clique size is shown to highly vary between 2 and 21 (Fig. 5.17). The diversity graphs of *acetate* and *glycolate* due to their star topology have maximum clique size equal to 2 (edge). The carbon source *L- asparagine*, on the other hand, generates a diversity graph of maximum clique size equal to 21.

The number of the metabolites that can be by-produced during the metabolism of the main carbon source plays an important role in the metabolic diversity that can emerge in the

system. The metabolites disturb the homogeneity of the growth medium and provide ecological opportunities (niches) for new phenotypes. The relation between the number of metabolites and the maximum metabolic variability is shown in Fig. 5.17. The active metabolic pathways are coupled decreasing the number of the different metabolic patterns that can emerge.

5.3.3 Consistent metabolic behaviors across conditions

This part aims to reveal consistent patterns of metabolic behavior such as mutants of a consistently unique metabolic blueprint or a consistently common metabolic profile with respect to by-production across the various growth conditions. Whether the environmentally (in-) variant metabolic behavior is related to genetic perturbations on the conserved (or non-conserved) part of the metabolism across divergent bacterial species is also investigated. The origin of mutants with the potential to develop cross-feeding interactions and lead a population to polymorphism is thus revealed.

5.3.3.1 Novel phenotypes

The strength (or degree) centrality is a measure of the importance of a node in a graph. Important nodes in a diversity graph correspond to mutants of different metabolic capabilities with respect to by-production than most mutants. In the previous section it was shown that the diversity graphs are highly centralized and a set of nodes of considerably high strength (or degree) centrality exists in each graph. These structurally important mutants comprise the core of the potential bacterial communities and are mainly responsible for the diversity and complexity that can potentially emerge in a bacterial population from single-gene deletions. In the following nodes of strength (or degree) centrality above a threshold of 0.5 are considered important in the weighted and the binary diversity graph representations respectively.

TABLE 5.8
METABOLIC INFORMATION OF THE CONSISTENTLY METABOLICALLY-UNIQUE MUTANTS OF *E. COLI*

Gene	Metabolic reactions	Metabolic sub-systems	ERI
‘b2276’	'NADH dehydrogenase ubiquinone 8 35 protons ' 'NADH dehydrogenase menaquinone 8 2 protons ' 'NADH dehydrogenase demethylmenaquinone 8 28 protons '	'Oxidative Phosphorylation'	0.59
‘b3731’	'ATP synthase four protons for one ATP '	'Oxidative Phosphorylation'	0.78
‘b2779’	‘enolase’	'Glycolysis-Gluconeogenesis'	0.97
‘b3236’	'malate dehydrogenase'	'Citric Acid Cycle'	0.81
‘b0116’	'2 Oxogluterate dehydrogenase' 'Glycine Cleavage System' 'pyruvate dehydrogenase'	'Citric Acid Cycle' 'Folate Metabolism' 'Glycolysis-Gluconeogenesis'	0.84
‘b2926’	'phosphoglycerate kinase'	'Glycolysis-Gluconeogenesis'	0.97
‘b0721’	'succinate dehydrogenase'	'Citric Acid Cycle' 'Oxidative Phosphorylation'	0.28
‘b0114’	'pyruvate dehydrogenase'	'Glycolysis-Gluconeogenesis'	0.38
‘b3956’	'phosphoenolpyruvate carboxylase'	'Anaplerotic reactions'	0.34
‘b2551’	'D alanine transaminase' 'alanine transaminase' 'glycine hydroxymethyltransferase' 'Threonine Aldolase'	'Cofactor and Prosthetic Group Biosynthesis' 'Cofactor and Prosthetic Group Biosynthesis' 'Glycine and Serine Metabolism' 'Threonine and Lysine Metabolism'	0.97
‘b3919’	'triose phosphate isomerase'	'Glycolysis-Gluconeogenesis'	0.94
‘b0529’	'methenyltetrahydrofolate cyclohydrolase' 'methylenetetrahydrofolate dehydrogenase NADP'	'Folate Metabolism' 'Folate Metabolism'	1.00

It is observed that a subset of these structurally important nodes-mutants consistently appears in most of the examined growth conditions independently on the graph representation (weighted or binary) that they have derived. In Figure 5.18 (Left) it is shown the appearance of each important mutant in the growth conditions where the corresponding deleted gene names the mutant. The metabolic reactions as well as the metabolic subsystem in which each of the corresponding deleted gene of each of the most frequently appeared mutants participates in is shown in Table 5.8.

5.3.3.2 Common phenotypes

Mutants of similar (common) metabolic capabilities with each other are characterized by low strength centrality in a diversity graph. Specifically, those mutants that fall in the lowest bin of the strength centrality distribution (see 5.3.2.2) are considered here as metabolically redundant. As shown in Figure 5.18 (Right), most of the redundant mutants are environmental-invariant, which means that the metabolic capabilities of the cell with respect to by-production are highly robust or redundant to the specific genetic perturbations and the examined growth conditions. The metabolic redundancy and the metabolic variability with respect to by-production are the mirror of one another and this property is also reflected on the conservation values of the genes responsible for each metabolic behavior.

5.3.3.3 The Evolutionary trait of specific phenotypes

A single-gene deletion can produce a phenotype with unique metabolic capabilities relative to other single-gene deletions in a given growth condition. Intuitively, one would expect that the most frequently-appearing unique mutants over the examined growth conditions might reflect deletions of environmental-invariant and essential genes that might also be evolutionary conserved over different organisms. This rational seems controversial to the evolution of polymorphism. If a gene appears to be evolutionary conserved the mutant that derives from its deletion is expected to be evolutionary unstable. However, it suggests first that the evolved robustness and vulnerability of the metabolic systems are reflected on the metabolic diversity graphs and its properties and second, that the two acting processes towards stabilizing either the monomorphic or the polymorphic state are antagonistic and that among all the potential communities probably only those that mainly consist of mutants that are environmental-specific eventually evolve. In the following it is shown that environmental-invariant mutants showing unique phenotypes relative to other mutants mainly correspond to deletions of conserved gene whereas consistently redundant mutants are mostly derived from deletions of non-highly conserved genes.

The ERI values of the corresponding deleted genes of each of the most frequently appeared mutants are shown in Table 5.8. Most of the environmental-invariant mutants (8 over 12) correspond to highly conserved gene deletions (ERI value above 0.7). To compare the environmental-invariant with the environmental-specific mutants with respect to the ERI values of the corresponding genes, a resolution step of 10 mutants is applied and the number of highly central mutants of which their corresponding deleted genes have ERI value above 0.7 is then determined. As shown in Fig. 5.18 (Left), it is observed that most of the consistently appearing central mutants (60% approximately) are related to the knockout of a highly conserved gene in contrast to the environmental-specific mutants of which approximately only the 30% correspond to highly conserved genes. The correlation coefficient between the gene conservation density and the frequency of appearance of the corresponding mutants equals to 0.6826 (p-value < 0.0025).

On the other hand, it is expected that the most frequently-appearing redundant mutants correspond to less evolutionary conserved genes with respect to the environmental-specific, since the metabolic capabilities of the cell at least with respect to by-production remain mostly unaffected by their deletion. Nevertheless, the environmental-specific redundant

mutants should have a non-redundant role in other growth conditions therefore one would expect that in that case their corresponding genes are more conserved. It is observed (Figure 5.18 - Right) that 50-90% of genes involved in the environmental-specific mutants are evolutionary conserved, whereas only 20-30% of genes are conserved in the environmental-invariant redundant mutants. The correlation coefficient between the gene conservation density and the frequency of appearance of the corresponding metabolically redundant mutants equals to -0.8224 (p-value < 10^{-10}).

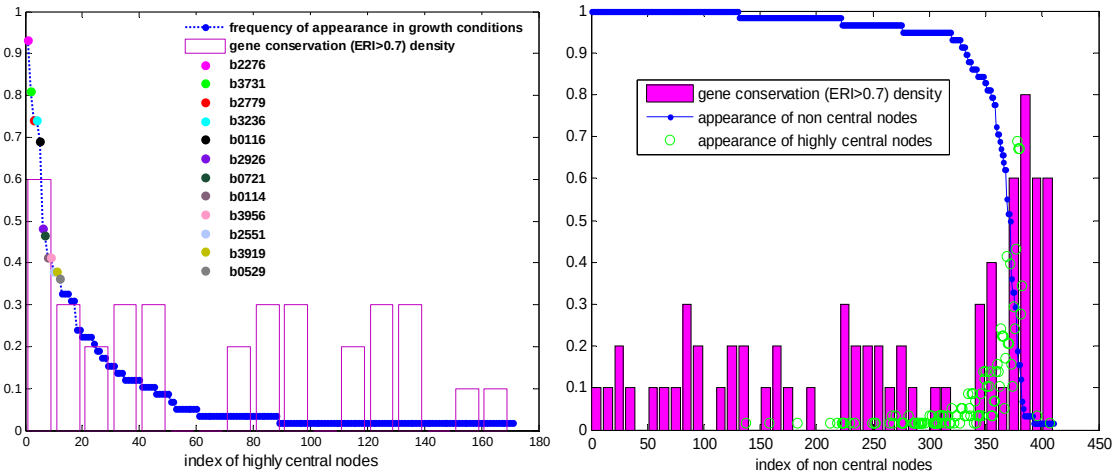


Figure 5.18 (Left) The frequency of appearance of each highly central mutant over the 58 different carbon-source conditions (blue line). The most frequently-appearing highly central mutants are highlighted. (Right) The frequency of appearance of each redundant mutant is shown. Redundant nodes in some growth conditions can be central mutants in other conditions. Their corresponding appearance is shown with green color. The gene conservation density, which expresses the number of evolutionary conserved genes of ERI value above 0.7 per every 10 highly central mutants, is also shown in the two figures.

5.3.3.4 Common by-products

Secondary metabolites are produced during the metabolism of the main carbon source that is provided for growth and are potentially utilized by the cells when the main source is exhausted in the growth medium. These metabolites comprise the sources of the metabolic variability and ecological opportunity allowing metabolic interactions to take place and polymorphism to evolve. *Formic acid*, *glycolate* and *acetate* are the most frequently appearing byproducts (Fig. 5.19) for the growth condition we have examined that can be observed if the system is genetically perturbed under single-gene knockouts (KO).

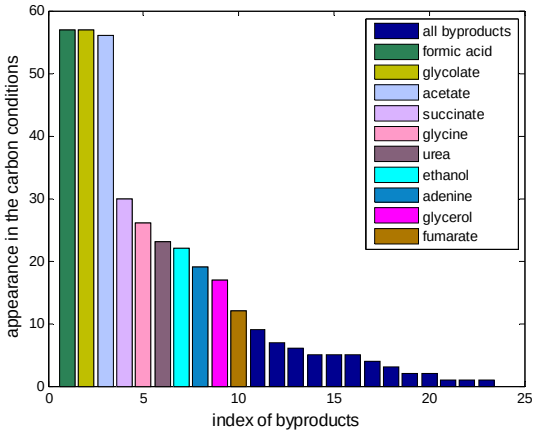


Figure 5.19 23 different metabolites were observed as by-products across the different carbon conditions under the applied genetic perturbations. The number of the growth conditions where each of these byproducts is present is shown. The most frequently-appearing metabolites are highlighted.

It was previously seen (Table 5.4) that when *glycolate* and *acetate* are used as a main source, only one mutant is observed to be metabolically different from all other mutants. Interestingly, this central mutant is actually the only mutant that is capable of producing *acetate* when *glycolate* is the main source (gene KO: 'b2276') and of producing *glycolate* when *acetate* is provided for growth (gene KO: 'b2978'). The diversity graph of *formic acid* is not provided because either the dynamic Flux Balance model we applied or the regulated model that incorporates transcriptional regulation in the Flux Balance model [200] are capable of simulating growth on this carbon source.

5.3.4 Conclusions based on Structural Analysis

The properties of the diversity graphs allow a global description of the metabolic behavior of the organism in different growth conditions under single-gene knockouts. Most single-gene deletions result in phenotypes that are similar to each other and to the wild-type cell with respect to by-production, reflecting the generally evolved metabolic robustness that characterizes most cellular systems. Nevertheless, a few mutants of novel metabolic capabilities exist and appear as hubs-highly central nodes in the diversity graphs, since they are metabolically different from most of the mutants of the graph.

The metabolic networks are inherently robust to genetic perturbation and environmental changes. Genome-scale deletion phenotype data for the bacterium *E. coli* [39] have shown that 83% of the examined protein-encoding genes (~87% of the total) were dispensable for cellular growth in rich media where indispensable is considered a gene that is either essential for cell viability or necessary for maintaining vigorous growth. The diversity graphs reflect this genetic robustness and show that the underlying plasticity mechanisms the cell invokes in order to compensate gene deletions rarely include rewiring of the metabolic pathways towards the production of exchange metabolites.

Furthermore, when the conservation value of the deleted gene is given as property to each node, a disassortative mixing is observed indicating a tendency of potential cross-feeding interactions to be developed between mutants derived from the deletion of a highly conserved gene and mutants derived from the deletion of a less conserved gene. This suggests that when the cell loses an essential, non-lethal, highly conserved gene then it can metabolically interact with another mutant derived from deletion of a non-conserved gene and get rescued from extinction.

The connectivity between the hubs is mostly high, forming the only clustered area in the graph. The central mutants of the diversity graphs comprise the core of the potential bacterial communities and are mainly responsible for the diversity and complexity that can potentially emerge in a bacterial population. Environmentally invariant mutants, which consistently appear as either highly different (central) or highly redundant (non-central) in the diversity graphs, have been identified as they might play an essential role in the evolution of the metabolic diversity in simple environments. It is observed that most environmental-invariant, highly central mutants correspond to deletions of highly conserved genes, which suggests that the two acting processes towards stabilizing either the monomorphic or the polymorphic state are antagonistic and that among all the potential communities probably only those that mainly consist of mutants that are environmental-specific eventually evolve.

The consequences of the particular structure that characterizes the diversity graphs include high network centrality, disassortativity on the strength centrality, a high mean clustering coefficient and dependence between strength centrality and clustering coefficient. Nevertheless differences in the graph properties over the different growth conditions exist. *Glycolate* and *acetate* generate diversity maps of star topology, which means that they are highly robust to genetic perturbations and only one mutant appears to be metabolically different with respect to by-production from all the others. In the case where *acetate* is used as a primal source the central mutant of the diversity graph is the only one capable of

producing *glycolate* and in the case where *glycolate* is used as the main carbon source the single central mutant that is found is actually the only one capable of by-producing *acetate*. It is observed that *glycolate* and *acetate* comprise the most common byproducts over the single-carbon cases we have examined. Growth on *adenosine* and *L tartaric acid* produce diversity graphs of comparatively low network centrality indicating growth conditions of higher metabolic variability. Metabolic pathways are coupled in a way that limits the number of potential divergent metabolic patterns. The maximum clique size that can be found in the diversity graph reflects the actual number of the different metabolic phenotypes and thus can better quantify the upper bound of the potential diversity that can emerge in a population in a given condition. *Acetate* and *glycolate* have maximum clique size equal to 2, whereas *L asparagine* exhibits 21 divergent mutants with respect to their metabolic capabilities as the maximum size.

5.4 Growth simulations of strain communities

As shown previously the diversity graphs can be used in order to identify strain communities consisting of genetically and metabolically different strains with the potential to develop cross-feeding interactions. Metabolic variability with respect to by-production disturbs the homogeneity of the medium allowing strains to interact with each other and novel phenotypes to be observed. The potential strain communities correspond to the cliques that exist in a diversity graph. To identify all the cliques, the binary and structurally compressed representation of the unconstrained diversity graph is used. However, not all the structurally identical mutants have the same growth characteristics and metabolic capabilities. Therefore the functional representation is used to determine the set of growth simulations that are evaluated. The growths of all the potential strain communities found in each diversity graph (functional representation) have been simulated for a subset of carbons sources including *glycolate*, *acetate*, *glycine*, *glucose, pyruvate* and *melibiose.* The first part has exhaustively investigated all the wild-type-mutant pairs under all carbon conditions. This part investigates coexistences of any potential size using the multi-competitor growth model.

5.4.1 Functional Analysis

In the Appendix it is mathematically proved that the conservation of mass does not allow group benefit to emerge from two different populations that compete for a primal limited source independently of their initial frequency or the amount of the resource. Nevertheless, other sources of heterogeneity such as by-production utilization and cross-feeding are shown that can lead a mixed population to supreme group performance when conditions that depend on the growth characteristics of the competitors and their initial frequency are satisfied. The growth conditions *glycolate, glucose, pyruvate* and *melibiose* are analytically presented in the following. The strain communities are studied with respect to their growth benefit, the involved cross-feeding interactions and the predictability of their growth performance (see 4.5).

A. Aerobic growth on *glycolate*

Single-growth characteristics

Growth on *glycolate* produces a diversity graph of star topology with only one mutant - the mutant of the deleted gene b2276- to be different from the rest mutants of the graph (including the wild-type cell). Because of this topology, the potential bacterial communities exclusively consist of mutant pairs of which the mutant of b2276 is omnipresent. This particular mutant is the only mutant capable of by-producing *acetate* and the only one incapable of producing *formic acid* contrary to the rest mutants (Fig 5.20). *Acetate* and *formic acid* are actually the two metabolites that are observed to be produced by the system. The gene b2276 exhibits unique metabolic properties in other growth conditions as well (Table

5.8). When deleted the cell exhibits poor growth performance (~20% reduction) compared to the growth performance of the wild-type cell in the same initial growth conditions. From all single-gene knockout mutants simulated to grow independently on limited *glycolate* no mutant is observed to be capable of combining maximum ability to metabolize *glycolate* with maximum ability to metabolize *acetate*, which is an essential by-product for growth. Furthermore, no mutant is observed to perform better than the wild-type cell (Fig. 5.20).

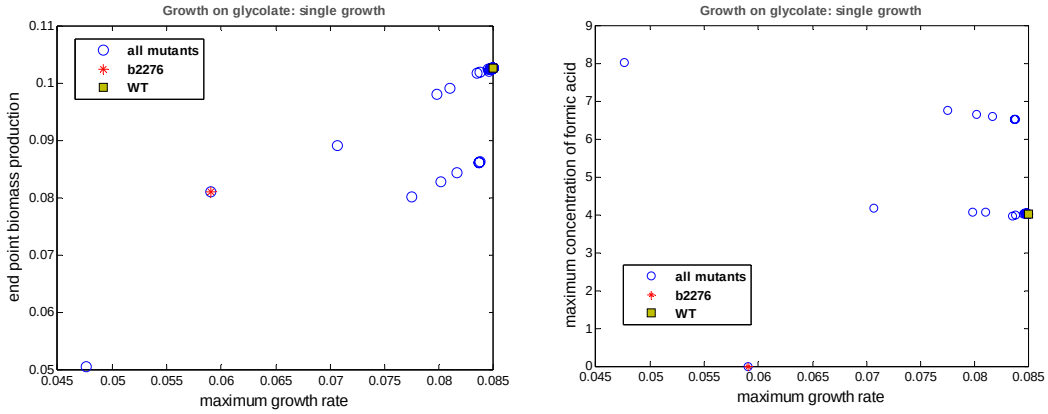


Figure 5.20 (Left) The maximum growth rate with respect to the end-point biomass concentration for each viable mutant growing independently on limited *glycolate*. (Right) All the mutants apart from the b2276 produce *formic acid*. The maximum growth rate with respect to the maximum concentration of *formic acid* is shown. The wild-type cell and the mutant b2276 are highlighted.

Two-competitor growth

The coexistence of the mutant of the deleted gene b2276 with any other mutant is observed to be beneficial, which means that the endpoint biomass concentration of the group is greater than the endpoint biomass concentration of each involved mutant growing independently as a monoculture (positive relative benefit). Absolute benefit is observed as well, which indicates that there are mutant coexistences able to better exploit the given environment than any single-gene knocked out mutant growing independently in the same initial conditions. As already shown in section 5.2.1, the co-growth of the wild-type and the mutant b2276 is an example of absolute benefit. The mutant of the deleted gene b2276 seems to have an altruistic behavior in its communities because of the *acetate* it provides to the growth medium. Nevertheless, it is observed to exploit the *formic acid* that is provided in the growth medium by the partner mutant increasing its growth rate (Fig. 5.21).

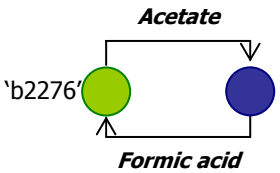


Figure 5.21 The cross-feeding interactions involved during the co-growth of a mutant with the mutant 'b2276' on limited *glycolate* lead the coexistence to supreme growth.

A metabolically interacting mutant pair consisting of the mutants derived from the deletion of the gene b2276 and the gene b3708 respectively is analytically described in the following. The flux rate time profiles (Fig. 5.22) verifies the exchange of the two products *acetate* and *formic acid* between them. Each novel metabolite is consumed simultaneously with the primal source. It is also observed that the growth rate of the mutant of b2276 is increased by the presence of *formic acid* in the growth medium and that also the growth rate of the mutant of b3708 is beneficially affected by the presence of the novel metabolite *acetate* in the growth medium (Fig. 5.23). Simulations of the coexistence between these two mutants in different initial population ratios were also performed. It is observed that when the mutant of b2276 is

present at a higher initial ratio in the population the more is the benefit of the coexistence (Figure 5.24). Maximum group performance is achieved with an initial population composition of 1:9 for b3708:b2276.

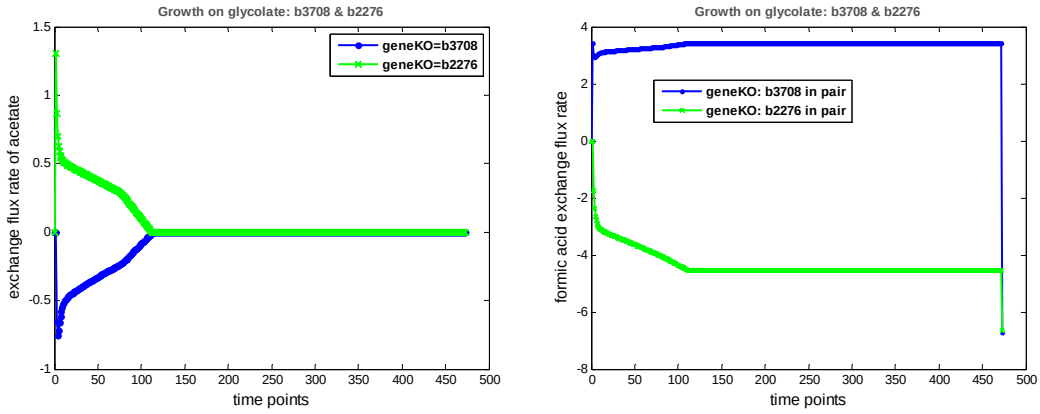


Figure 5.22 Two mutants, the b2276 and the b3708, co-grow on limited *glycolate.* The flux rate time profiles of two metabolites *acetate* (Left) and *formic acid* (Right) are shown for each mutant under competition.

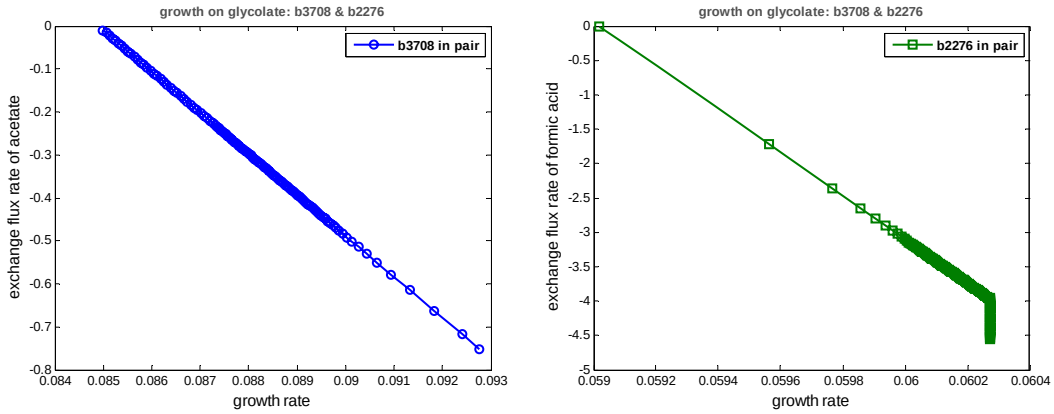


Figure 5.23 (Left) The growth rate of the mutant b3708 is increased as the concentration of *acetate* increases in the medium and consequently its uptake flux. (Right) The growth rate of the mutant b2276 is increased as the concentration of *formic acid* increases in the common medium and consequently its uptake flux.

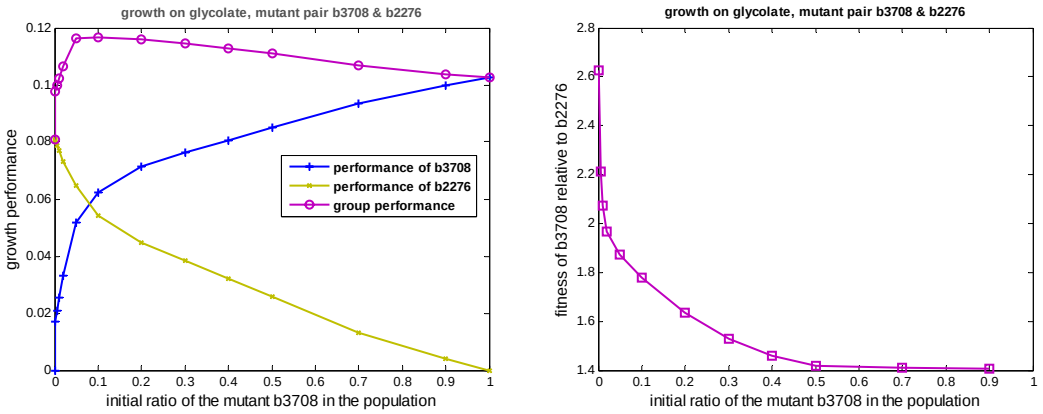


Figure 5.24 The growth performance of the mutants of b3708 and b2276 as they co-living on limited *glycolate* for various initial population ratios. The group performance increases as the mutant of b2276 is (initially) present in higher ratio in the population than the mutant b2276. Maximum group

performance is observed when the two mutants exhibit equal final frequency in the population, which is achieved with an initial population composition of 1:9 for b3708:b2276.

B. Aerobic growth on *pyruvate*

Single-growth characteristics

The set of the metabolites that are observed to be by-produced by the mutants during their independent growth on limited *pyruvate* include: *acetate*, *formic acid*, *glycine* and *glycolate.* The maximum clique size that exists in the diversity graph of *pyruvate* is 6, whereas the number of the highly central nodes that correspond to mutants of highly different metabolic capabilities with respect to by-production is 7. Among all single-gene knockout mutants, the mutant derived from the deletion of the gene b3403 is observed to exhibit the maximum growth performance, which is 4.7% increased with respect to the performance of the wild-type cell population (Fig. 5.25). The gene b3403 is not highly conserved among different bacterial species (ERI equals to 0.38), encodes the *phosphoenolpyruvate carboxykinase* and participates in the *anaplerotic* reactions of the *E. coli* metabolism. The underlying reason for its high growth performance is the efficient strategy which follows during its growth even if it costs to the cell a lower growth rate with respect to the wild type during the metabolism of *pyruvate*. Specifically, this mutant exhibits a relatively high by-production of *acetate*, which uses as secondary resource when *pyruvate* is exhausted (Fig. 5.25).

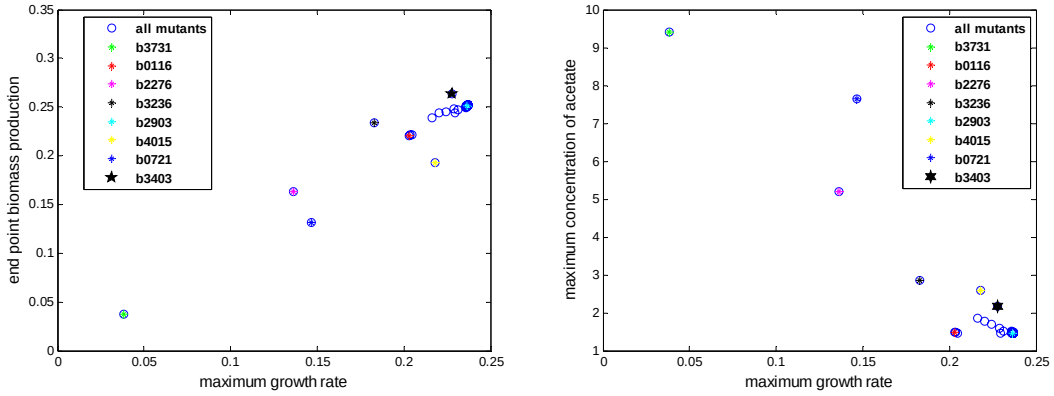


Figure 5.25 (Left) The maximum growth rate with respect to the end-point biomass concentration for each viable mutant growing independently on limited *pyruvate*. (Right) The maximum growth rate with respect to the maximum concentration of *acetate* is shown. The highly central mutants are highlighted. The most efficient single-growth mutant (b3403) is also depicted.

Two-competitor growth

The absolute and relative benefits of all the examined mutant pairs when co-grow on limited *pyruvate* are shown sorted in Fig. 26. The highly central mutant involved in each pair is also shown. It can be observed that the mutant b3236 and b0116 exhibit consistently negative absolute benefit in any pair, whereas the mutant b3236 exhibits in addition consistent negative relative benefit indicating its inefficient contribution in co-growth.

On the other hand, when the environmental-invariant mutant of the deleted gene b3731 (Table 5.8) coexists with the mutant of the deleted gene b3403, which exhibits the best single growth performance on *pyruvate*, they beneficially interact (absolute and relative benefit equals to 0.8%). The mutant of the deleted gene b3731 exhibits the lowest growth rate of all the viable mutants (Fig. 5.25) that can be derived from single-gene knockouts that grow on limited *pyruvate*. This mutant is also observed to exhibit the maximum *acetate* production (Fig. 5.25). However, it is not capable of consuming the *acetate* it produces. The

interaction concerns the exchange of *acetate* and *formic acid* from the b3731 to the b3403 because the last is the only mutant that is capable of consuming these metabolites. Furthermore, the interaction concerns the exchange of *glycolate* from the mutant of b3403 to the b3731 as can be seen in the exchange flux rate profiles (Fig. 5.27). It is observed that the presence of *glycolate* increases the growth rate of the mutant of b3731. The growth rate depends linearly on the uptake flux rate of *glycolate* (Fig. 5.27). The absolute value of the uptake rate increases as the amount of *glycolate* increases in the growth medium since the availability of the metabolite directly shapes the corresponding capacity flux bounds.

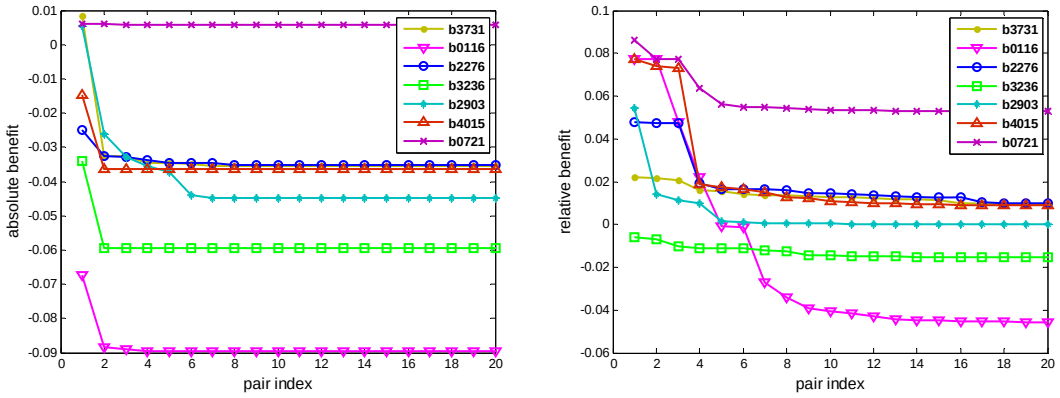


Figure 5.26 Each mutant pair corresponds to the potential interaction between a highly central mutant and another mutant in the diversity graph. (Left) The absolute benefit of the mutant pairs when co-grow on limited *pyruvate*. (Right) The relative benefit of the mutant pairs.

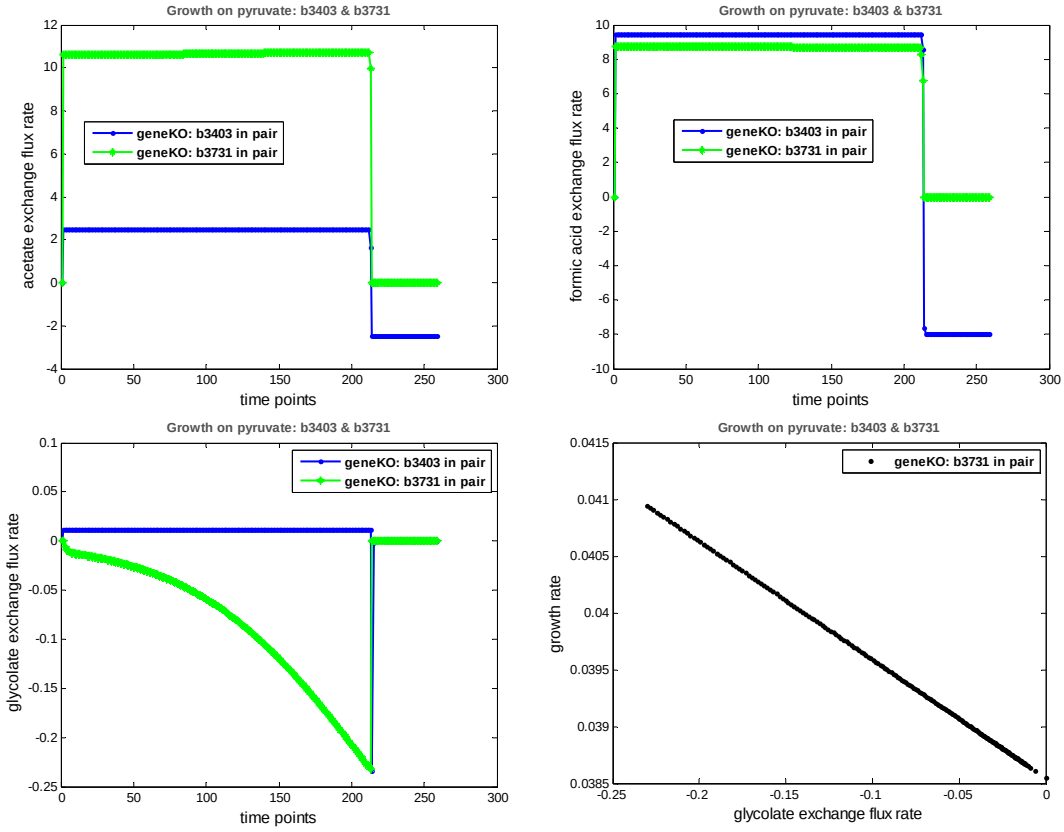


Figure 5.27 Two mutants, the b3403 and the b3731, co-grow on limited *pyruvate.* The flux rate time profiles of the metabolites *acetate* (Top-Left), *formic acid* (Top-Right) and *glycolate* (Bottom-Left) are shown for each mutant under competition. The growth rate of the mutant b3731 is increased as the concentration of *glycolate* increases in the medium and consequently its uptake flux.

The growth performance of the most beneficial mutant pair (among all the pairs) has been also explored for different initial ratios of the mutants in the population. It is observed (Fig. 5.28 - Right) that maximum growth benefit equals to 0.8% is achieved when the two mutants are initially in equal frequency in the population. On the other hand, another pair, which consists of the mutants of the deleted gene b3403 and b0721 respectively, is observed to exhibit beneficial growth performance of absolute benefit equal to 1.1% at the initial ratio of 0.9:0.1 of b3403:b0721 (Fig. 5.28 – Left). This example is presented in order to underline the importance of the initial population ratios to the group performance of the communities and that exploring equal initial frequencies covers a sub-space of potential beneficial communities.

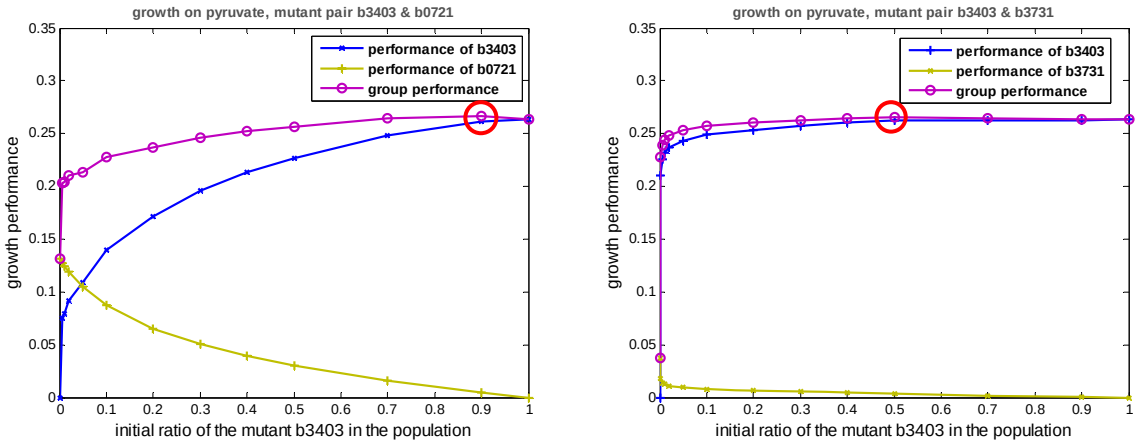


Figure 5.28 (Left) The growth performance of the mutants of b3403 and b0721 as they co-grow on limited *pyruvate* for various initial population ratios. The two mutants, b3403 and b0721 in equal initial frequency exhibit no relative benefit, since the growth performance of the group is below the performance of the mutant of b3403. However, as the population of the mutant of b3403 increases a benefit is observed (indicated with a red circle). (Right) The growth performance of the mutants of b3403 and b3731 as they co-grow on limited *pyruvate* for various initial population ratios. This mutant pair is beneficial when the mutants are initially in equal frequencies.

Multi-competitor growth

The growth of all the strain communities of each possible size has been simulated. Absolute and relative benefits of positive values are observed in several cases (Fig. 5.29). Analysis of these strain compositions based on the flux and growth rate time profiles reveals that in all cases metabolic interactions are involved. Furthermore, when a mutant provides a novel metabolite, the mutant that is capable of catabolizing it also increases its growth rate. The involved metabolic interactions of the most efficient strain triplet and the most efficient strain pair mentioned previously are shown in Figure 5.30.

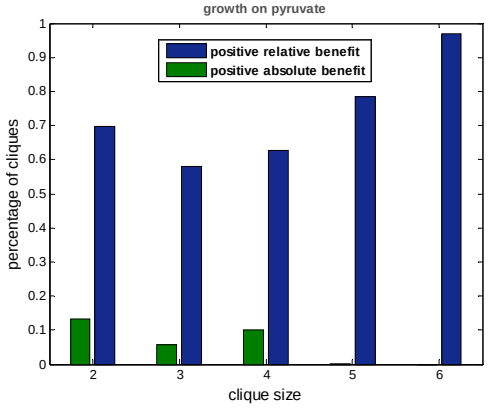


Figure 5.29 The percentage of strain communities exhibiting either relative or absolute benefit is shown.

A subset of the highly central mutants and specifically those that corresponds to the knockout of the gene b0721, b4015, b2276 and b3731, is observed to play a beneficial role (relative benefit >0) in most of the pair interactions in which each of them participates (Fig. 5.26). Actually these mutants produce metabolites such as *acetate* and *formic acid*, which however are incapable of consuming (exception is the mutant of b2276, which is observed to consume the *formic acid* it produces). In their metabolic interactions, these mutants play the role of a provider of the specific nutrients, which are proved essential for the growth of the other members of the community. As expected when these mutants coexist with each other and since none of them is capable of consuming the available nutrients no benefit is observed. The unexplored mutant pairs are incapable to develop cross-feeding interactions but have been introduced in the diversity graph as unconstrained edges (see 4.2.2) because in larger communities are observed to play an important and novel role since they allow another strain from the community to (beneficially) exploit the available products.

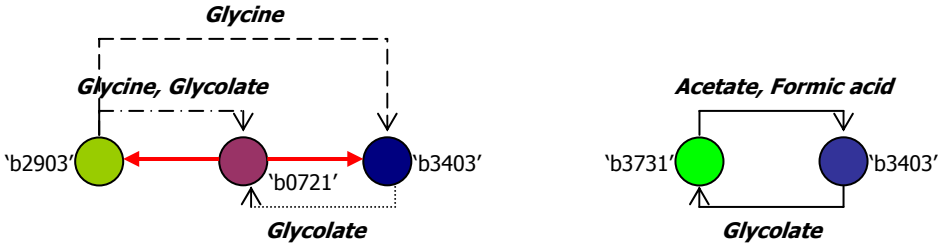


Figure 5.30 (Left) The strain triplet of b2903, b3403 and b0721 exhibits the maximum absolute benefit (1.4%) over all the simulated communities. The arrows show the metabolites that are exchanged during growth on limited *pyruvate*. The red arrows correspond to the exchange of *acetate* and *formic acid,* metabolites which only the mutant of b0721 is incapable of consuming. (Right) The strain pair b3731 and b3403 exhibits absolute and relative benefit equal to 0.8% when co-grows on limited *pyruvate*. The metabolites that are exchanged between them are shown.

Each community is comprised of smaller sub-communities, whereas the smallest size communities correspond to pairs. In Figure 5.31 (Left), it is shown how the unexplored mutant pairs introduce non-linearities in the growth performance predictions of the larger communities when prediction is based on knowledge of the performance of the constituent pairs. Nevertheless, when the constituent triplets are considered (introduction of effective weights-see 4.5.2.1) for the prediction of larger communities instead of the pair-wise interactions, the correlation coefficient is significantly (Table 5.9) high (p-value = 0). All p-values in Table 5.9 equal to 0 unless it is differently indicated. Figure 5.31 (right) shows an example of the relation between the simulated performance and its prediction, which is based on pairs (depicted with a blue color) and on triplets (depicted with a cyan color).

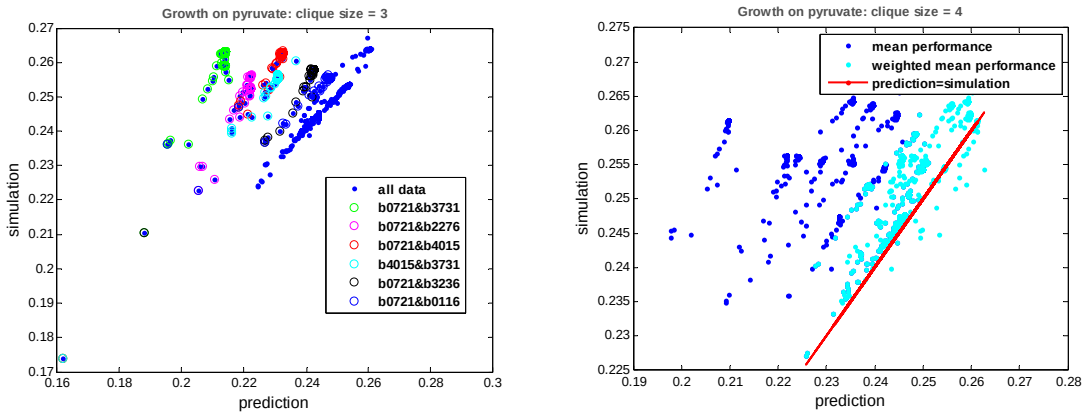


Figure 5.31 (Left) The simulated growth performance of all the triplets that are identified in the diversity graph of *pyruvate* with respect to the mean performance of the pair-wise interactions

(prediction) is shown. A subset of the edges between the highly central mutants as participate in the triplets is also shown. (Right) The simulated growth performance of all the cliques of size 4 with respect to the predicted performance, which is estimated by the mean performance of the performances of the pair-wise interactions that each cliques consists of (blue color) or by the weighted mean performance where the effective weights are introduced (cyan color) for certain interactions.

Table 5.9: Growth on *pyruvate* – Correlation coefficient between predicted and simulated group performance

Clique size:	3	4	5	6
R (based on pairs):	-0.0328 (p-value=0.066)	-0.2404	-0.3585	-0.6206
R (based on triplets):	0.8325	0.8278	0.8320	0.9529
Number of cliques:	3130	3487	1822	364

C. Aerobic growth on *glucose*

Multi-competitor growth

The growth of all the different mutant compositions has been evaluated. Absolute benefit is not observed (Fig. 5.32) in any clique of any size, which means that no bacterial community consisting of metabolically different, single-gene knockout mutants has been found to exploit better the given environment than the monoclonal populations. This observation implies that within the pool of single-gene knockouts, certain mutants such as the mutant cells of the deleted gene b0114 or the wild-type cells are observed to be capable of best exploiting the given environment when growing as monocultures. The mutant of the deleted gene b0114 exhibits end-point biomass, which is 0.2% higher than the performance of the wild-type in the same initial growth conditions. This mutant is observed to produce relatively less *acetate*, while it is not capable of by-producing *glycolate*. On the other hand, several mutant compositions are observed to exhibit positive relative benefit. The environmental-invariant mutant of the deleted gene b0721 (Table 5.8) produces *acetate*, *glycolate* and *formic acid* when grows on limited *glucose,* metabolites, which however it is not capable of consuming. In communities, this peculiar mutant thus plays the role of a provider of essential metabolites for growth having a purely altruistic behavior.

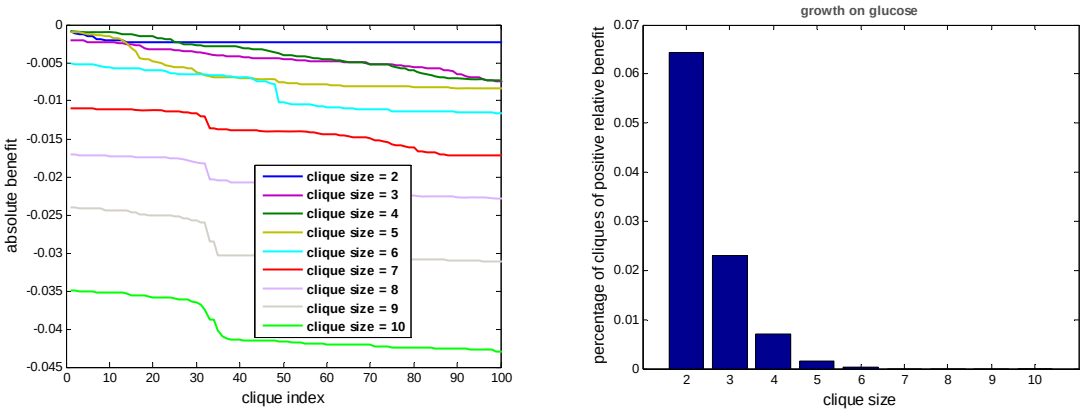


Figure 5.32 (Left) Absolute benefit is not observed (negative value) in any clique of any size. (Right) The percentage of strain communities exhibiting relative benefit is shown.

It is observed that knowledge of the growth performance of the mutant pairs is sufficient to accurately predict the group performance of larger communities (Table 5.10). Significant (p-value = 0) high correlation coefficient is observed between the predicted and the simulated growth performances. Nevertheless, knowledge of the mutant triplets improves the predictability when the communities are of size large enough (Table 5.10) as the significant (p-value = 0) high correlation coefficient show. All p-values in Table 5.19 equal to 0. In

Figure 5.33 it is shown that the maximum divergence from the mean performance corresponds to cliques where unconstrained edges are present.

Table 5.10: Growth on *glucose* – Correlation coefficient between predicted and simulated group performance

Clique size:	3	4	5	6	7	8	9	10
R (based on pairs):	0.8486	0.8208	0.8165	0.8062	0.8016	0.8061	0.8259	0.7540
R (based on triplets):	0.8799	0.9126	0.9283	0.9357	0.9323	0.9234	0.9145	0.9027
Number of data (cliques):	2087	6321	12229	15589	13077	6962	2136	288

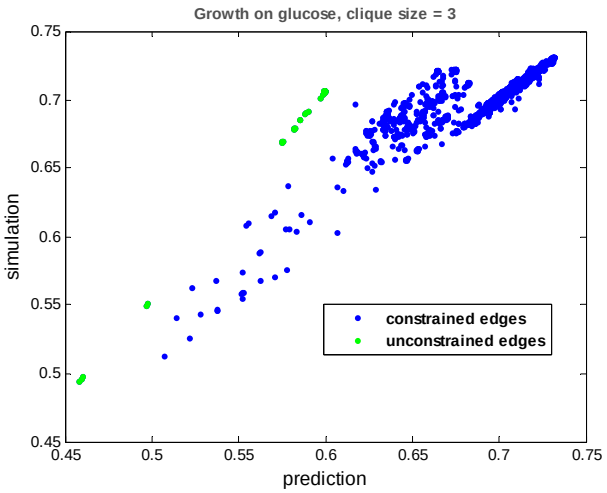


Figure 5.33 The simulated growth performance of all the triplets that are identified in the diversity graph of *glucose* with respect to the mean performance of the pair-wise interactions (prediction). The mutant triplets consisting of constrained edges are indicates with blue color. The complementary triplets that include unconstrained edges are shown with green color.

D. Aerobic growth on *melibiose*

Single-growth characteristics

During growth on limited *melibiose* a number of secondary metabolites are observed to be by-produced by the mutants such as *acetate*, *dihydroxyacetone*, *ethanol*, *formic acid*, *fumarate*, *glycerol*, *glycolate* and *succinate*. Among all single-gene knockout mutants a subset was found to perform better than the wild-type cell under the same initial growth conditions. Maximum single growth performance exhibits the mutant of the deleted gene b1602. This gene of ERI value equal to 0.44 encodes the *NAD transhydrogenase* and the *NAD P transhydrogenase* participating in the *Oxidative Phosphorylation* pathway. Its high growth efficiency lies in its ability to efficiently catabolize the *melibiose*, the main carbon source and its combined high efficiency to metabolize *ethanol* among other metabolites such as *acetate* and *formic acid*. This mutant is also a highly central mutant in the diversity graph of *melibiose*.

Multi-competitor growth

Absolute benefit is not observed in any clique of any size indicating that there isn't strain community capable of exploiting better the given environment than the mutant of the deleted gene b1602. Nevertheless, relative benefit is observed in a few communities (Fig. 5.34). The strain pair 'b3731' and 'b2779' exhibits the maximum relative benefit that is observed (~7%). The involved metabolic interactions of the beneficial strain pair are shown in Figure 5.34. The mutant, which is derived from the deletion of the gene 'b3731', is observed to produce the maximum concentration of *acetate* among all mutants. However, it is incapable of consuming the produced *acetate* and it is also incapable of consuming the formic acid that produces.

Nevertheless, it is observed to consume the ethanol it produces. The mutant 'b2779', on the other hand, is capable of exploiting the available *formic acid* and partly the *acetate* that are produced in excess in the medium. Thus, when the two strains coexist the mutant 'b3731' plays the role of a provider of essential products to the mutant 'b2779'. The strain triplet 'b2926', b2276' and 'b2779' is also observed to exhibit similar relative benefit.

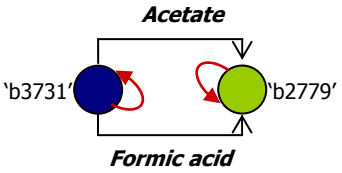


Figure 5.35 The arrows show the metabolites that are exchanged when the strain pair 'b3731' and 'b2779' is growing on limited *melibiose*. The mutant 'b3731' provides acetate and formic acid to the mutant 'b2779'. The red arrows correspond to *ethanol,* a metabolite which both mutants are capable of metabolizing.

It is observed that knowledge of the growth performance of the mutant pairs is sufficient to accurately predict the group performance of larger communities. Significant (p-value = 0) high correlation coefficient is observed between the predicted and the simulated growth performances (Table 5.11). However, unexploited pairs such as the mutants of b3731 and b2926 and the mutants of b3919 and b2926, which are incapable of fully exploiting the given environment and specifically of consuming *acetate* and *formic acid*, essential products for growth, exist causing divergent of the prediction from the simulated performance. Thus, an improvement in the correlation coefficient can be observed when prediction is evaluated based on simulated triplets (Table 5.11). All p-values in Table 5.11 equal to 0.

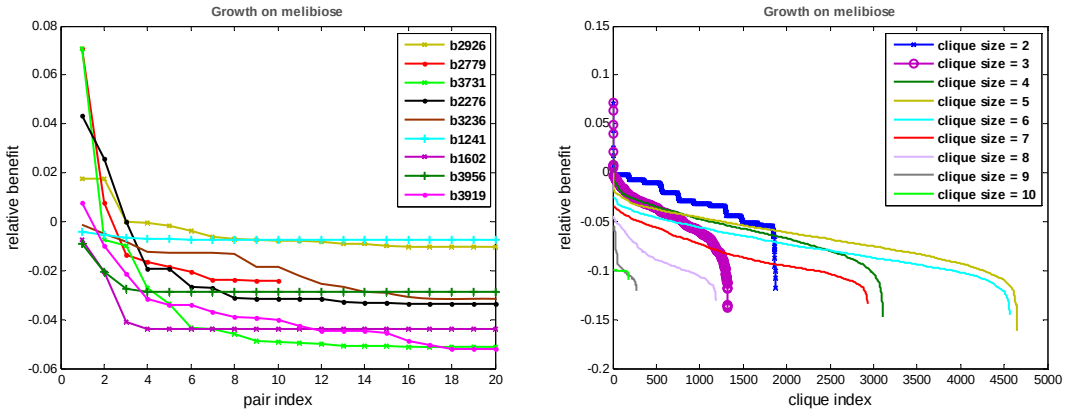


Figure 5.34 (Left) Relative benefit of the mutant pairs under growth on limited *melibiose*. The pairs of each highly central mutant are shown. (Left) Relative benefit of each clique. Relative benefit is not observed in any clique of size greater than five.

Table 5.11: Growth on *melibiose*– Correlation coefficient between predicted and simulated group performance

Clique size:	3	4	5	6	7	8	9	10
R (based on pairs):	0.8878	0.8854	0.8573	0.8916	0.9092	0.9190	0.9350	0.9034
R (based on triplets):	0.8794	0.8901	0.8789	0.9166	0.9217	0.9290	0.9178	0.8387
Number of data (cliques):	1322	3108	4648	4564	2934	1187	273	181

5.4.2 Conclusions based on Functional Analysis

The diversity graphs are reconstructed in order to reflect the potential of mutants to exchange products of their metabolism, which was defined as the pair-wise differences in the by-production efficiency of the mutants. The metabolic capabilities of each mutant are determined by their growth simulations as monoculture. The uptake capability of each mutant

to exploit the available metabolites could have been incorporated as information in the reconstruction of the potential communities to further restrict the search space of the possible mutant compositions. However, the monoculture growth simulations cannot provide direct information regarding a novel substrate that the presence of another mutant potentially provides in the growth medium neither can predict the effect of the dynamically shaped medium on the growth performance of the community. Whether the mutants actually interact with each other by exchanging nutrients can be observed through the flux rate time profiles of the exchange reactions when co-grow on the given environment, information that is provided by the multi-competitor FBA model. Furthermore, the relations between the metabolic variation that is met in a potential community with the actual metabolic interactions that take place and the implied efficient or inefficient exploitation of the given environment have been thoroughly explored.

The main conclusion is that beneficial metabolic interactions between the members of a community can lead the bacterial population to supreme growth without the application of any prior, common objective. As suggested mathematically (Appendix), neither absolute nor relative benefit (of positive value) is observed unless there are metabolic interactions between the members of a community. In several growth conditions, no mutant among all single-gene knockout mutants simulated to grow independently is observed to be capable of combining maximum ability to metabolize the main source with maximum ability to metabolize essential products of the metabolism, while a community of metabolically divergent strains is found to better exploit the given environment (absolute benefit). Bacterial communities that exhibit absolute benefit have been observed under growth on *glycolate* and *pyruvate*. The existence of interacting heterogeneous populations capable of better exploiting a given growth medium than monocultures indicates that in some growth conditions, the involved metabolic pathways are coupled in a way that a single optimal mutant is incapable of fully utilizing the environment. Similar findings have been observed in long-term evolutionary experiments with *E. coli* [19-22]. From the evolutionary perspective, the evolution of polymorphism against the domination of a fittest monoclonal population might be determined by the metabolic capabilities of the derived mutants that arise naturally by undirected processes and their potentiality to exploit the given environment in a way that results in an increase of their population.

Beneficial cross-feeding interactions can occur either directly or indirectly. The indirect metabolic interactions imply that a bacterial community can be beneficial even if not all of its pair-wise relations correspond to metabolic interactions where exchange of essential nutrients takes place, which verifies the important addition of the unconstrained edges in the diversity graph reconstruction. Furthermore, it is observed that beneficial metabolic interactions can be either bi-directional where the exchange of essential nutrients flows in both directions, so that both mutants exploit the newly shaped environment or they can be in one direction where only one benefits from the coexistence and the other plays the role of a mere provider, an altruist. However, metabolic interactions are also observed to take place without a beneficial outcome (negative relative benefit). Negative relative benefit corresponds to cases where the gain of utilizing the available nutrients that are novel or in excess as secondary resources (if metabolic interactions take place) is less than the loss of co-growing with less efficient individuals. It is observed that the presence of novel metabolites for a specific mutant in the growth medium may beneficially affect its growth rate and consequently its growth efficiency in the population. Parallel consumption of the novel metabolite with the primal source has been observed. Furthermore, as the amount of essential nutrients accumulates in the common environment, the upper bounds of the capacity flux constraints might be reached whereas a correlation between the exchange flux rate and the growth rate is observed below the maximum uptake rate.

The predictability of the growth performance of the communities based on knowledge of the performance of their smaller constituent sub-communities such as pairs or triplets is also investigated because it underlies how far the novelty with respect to by-production propagates in strain compositions altering the growth medium. It is observed that as long as

pairs of mutants of unexplored or unexploited metabolic capabilities are not present, the performance of the clique linearly depends on the mean performance of the pair-wise interactions as the high correlation coefficient values indicate for the cases of growth on *melibiose* and *glucose*. For growth on *pyruvate* however, several edges of unexplored metabolic capabilities participate in larger communities where their contribution is considerably affected. In that case the decomposition of the cliques is reduced to triplets instead of edges regarding the prediction of the growth performance. Probably as the composition of the communities becomes more complex consisting of more specialized mutants and allowing more obligatory relations to develop between them, the predictability of the growth performance of the communities from their simplest constituents to be lost.

Considering that the metabolic genes can have a wide distribution of essentiality in a given cell and across species and taking also into account that the potential communities by reconstruction have higher probability of consisting of mutants derived from highly conserved gene knockouts (section 5.3.3.3) giving them lower probability to evolve, the finding that the efficient communities (positive absolute benefit) do not consist of mutant compositions that in their majority correspond to deletions of highly conserved genes is important under the evolutionary perspective. Among the potential communities we have examined no absolute benefit was observed to correspond to compositions where more than one mutant has derived from the deletion of a highly conserved gene of ERI value greater than 0.6. Only few cases have been observed in communities of size 4, on growth on *pyruvate* to consist of mutants of which exactly two of them correspond to genes of ERI value in (0.7 0.8]. This study also suggests that if in a population a mutant arises by the knockout of an important gene that considerably alters the metabolic capabilities of the cell affecting part of the population, the cellular population is possible to continue functioning efficiently when metabolic interactions take place.

6. Conclusion

The evolution of cooperative strategies poses a challenge to the theory of natural selection, whereby only the fittest individuals survive. However, diversity rapidly emerges even in the simplest environments and the simplest organisms and microbial life does not comprise an exception. A variety of social behaviors involving communication and cooperation have been observed in microbes showing that they form socially interacting systems in the same manner as many other species on earth.

Long-term evolution experiments on monoclonal populations of the bacterium *Escherichia coli* growing in a homogeneous environment consisting of a single-limiting resource have verified the emergence and maintenance of more than a single strain in the population. In a homogeneous medium of a single-limiting resource, the competitive exclusion principle predicts that only one genotype the fittest will eventually be maintained in an asexual population. This scenario can change however when the growth of one genotype creates a new niche for another genotype as it excretes usable metabolites that can be served as secondary resources for growth. The availability of additional resources allows beneficial cross-feeding interactions between the members of the population to take place, which may underlie the stable co-existence of multiple strains.

In an attempt to improve our understanding of the evolution of the metabolic diversity in simple environment and the mechanisms supporting cooperative behaviors, this work goes a step further from single-cells; it develops the first genome-scale metabolic model capable of simulating a competitive life within cell communities, where different individuals co-grow, sense, shape and respond to a common, dynamic environment. Since bacteria have evolved to maximize their growth, the way the growth performance is shaped under competition has been thoroughly investigated. The metabolic model has provided insights about the mechanisms underlying the emergence of cooperative societies consisting of self-centered individuals, the formation of which allows immediate benefits to emerge both at the individual and the community level. It was analytically proved that in a simple and spatially homogeneous environment where by-production is not allowed to disturb the medium, competition for the primal source alone cannot lead a heterogeneous population to group benefit.

This work used the bacterium *E. coli* as a case study. The quantitative effect of the single-gene knockouts on the metabolic capabilities of the cell with respect to by-production has been computationally examined in order to predict strain communities with the potential to develop cross-feeding interactions. Gene inactivation or loss comprises a common strategy of the adaptation process of a bacterial population to a specific environment.

Based on the assumption that mutants of similar metabolic capabilities sharing a common growth environment do not contribute any new experience or metabolic interaction to the community contrary to mutants of different capabilities, a graph representation (diversity graph) has been developed in order to map the genetic to the metabolic variability with respect to by-production. This mapping is also supported by the evolutionary experiments,

which have shown that the metabolically interacting evolved strains considerably differ with each other with respect to their metabolic capabilities and gene expression patterns. The diversity graph reconstruction was built upon the genome-scale metabolic network reconstruction of *E. coli* as well as the constraints and assumptions of the existing dynamic FBA model. The diversity graphs were reconstructed for a variety of carbon sources. Specific graph properties were explored in order to provide an overall picture of the metabolic capabilities of the cell under specific genetic perturbations as well as to characterize and compare these networks in different growth conditions. Communities that have the potential to interact metabolically contained strains with different metabolic capabilities and corresponded to the cliques of the diversity graph. The growths of these communities were simulated in several growth conditions utilizing the developed genome-scale multi-competitor metabolic model.

This work shows that metabolic interactions are indispensable within strain communities in order to perform efficiently under conditions of resource competition. The existence of interacting heterogeneous populations capable of better exploiting a given growth medium than monocultures indicates that in some growth conditions, the involved metabolic pathways are coupled in a way that a single optimal mutant is incapable of fully utilizing the environment. Furthermore, it was observed that in several efficient cases co-growth provides immediate benefits to the competitors by increasing their growth rate.

Metabolic diversity was demonstrated to serve as the seed for the emergence of metabolic interactions within populations. It was observed that the evolved robustness and vulnerability of the metabolic systems across various single-carbon sources and across divergent species is reflected on the metabolic diversity graphs. Environmental-invariant important nodes correspond to mutants of essential gene deletions. The graph analysis suggested that the two acting processes towards stabilizing either the monomorphic or the polymorphic state are antagonistic and that among all the potentially interacting communities probably only those that mainly consist of mutants that are environmental-specific are likely to evolve.

Since the initial population frequency of the competitors as well as the amount of the main source that is initially provided for growth play an important role in the growth performance of the communities as they determine the partitioning of the resources, it is expected that the beneficial communities are not limited to our observations even within the given search space. As complexity increases and as environments become more complex than the homogeneous medium of a single-limiting resource that was explored in this study, the maintenance of diversity might prove far more beneficial for the systems involved. In a continuously changing, multi-tasking and complex real world with multiple demands, multi-cellular interacting systems might better exploit the environment than a monomorphic population.

The predictive accuracy of the proposed multi-competitor model depends on the accuracy of the genome-scale metabolic reconstructions and their reliable prediction of the transport fluxes under genetic perturbations and growth conditions. Furthermore, the cellular interacting system is usually decomposed into signaling, regulatory and metabolic network. However, these networks are not independent and an integration of all the networks into a complete genome-scale reconstruction of cellular functions at the level of individual chemical reactions remains to be established.

Different definitions of metabolic variability can also be considered in the graph reconstruction. Other sources of genetic to metabolic diversity beyond single gene knockouts, which have been studied in this work, such as multiple gene deletions or differential expression of certain genes are also possible to contribute as nodes to the graph reconstruction and as potential competitors in strain communities. Furthermore, the incorporation of spatial heterogeneity or signaling queues such as quorum sensing (able to sense the surrounding population) might provide the interacting system with other interesting capabilities. The metabolic diversity and interactions across different species can be also

analyzed under the framework proposed here as more genome-scale metabolic models being reconstructed. Apart from the capability of the proposed growth model to accommodate in its framework any number of different strains or species that communicate with each other through the environment, cells following different regulatory programs or cells with different objectives-roles as in multi-cellular organisms can be used as well.

Apart from the biological significance of this study, knowledge of the potential diversity that can emerge in a population might be important in other areas such as biotechnological applications where growth-efficient bacterial systems or waste-product depletion systems are searched or in human health when antibiotic drug-targets select bacterial enzymes. In a scenario where a drug affects part of the bacterial population altering its metabolic capabilities it might be the case that instead of suppressing the bacterial growth to enhance it through their metabolic interactions. The method presented in this work has many implications for research on the ecology of increasingly complex microbial communities in natural and engineered environments. Furthermore, the identification of heterogeneous bacterial cultures with superior desired properties might further exhibit a broad range of applications in metabolic engineering.

The publications that have resulted from this thesis so far include:

A. Book chapters

Eleftheria Tzamali, Panayiota Poirazi and Martin Reczko, "Methods for Dynamical Inference in Intracellular Networks", Bioinformatics for Systems Biology, Humana Press, 28, 541-561, 2008

Maria Manioudaki, **Eleftheria Tzamali**, Martin Reczko, Panayiota Poirazi, "Methods for structural inference and functional module identification in intracellular networks", Bioinformatics for Systems Biology, Humana Press, 27, 517-540, 2008

B. Conferences

E. Tzamali, P. Poirazi, I. G.Tollis, M. Reczko, "Computational identification of bacterial communities", *International Journal of Biological and Life Sciences*, 1(4):185-191, 2009

E. Tzamali, M. Reczko, "The benefit of cooperation: Identifying growth efficient interacting strains of Escherichia coli using metabolic flux balance models", *8th IEEE International conference on bioinformatics and bioengineering,* GREECE, 2008

C. Poster

E. Tzamali, P. Poirazi, I. G.Tollis, M. Reczko, "Computational study of the metabolic diversity of the bacterium *E. coll*", 2nd Annual Conference of the Greek National Initiative "Mikrobiokosmos", Athens, Greece, December 11-13, 2009

D. Journal

E. Tzamali, P. Poirazi, I. G.Tollis, M. Reczko, "Comparative Computational Analysis of the Metabolic Diversity of *E. coli* across Growth Conditions and Exploration of Growth-Efficient Polymorphic Communities", in preparation

References

1. Nowak MA, Sigmund K: **Evolutionary dynamics of biological games**. *Science* 2004, **303**(5659):793-799.
2. Pfeiffer T, Schuster S: **Game-theoretical approaches to studying the evolution of biochemical systems**. *Trends in biochemical sciences* 2005, **30**(1):20-25.
3. Clutton-Brock TH, O'Riain MJ, Brotherton PN, Gaynor D, Kansky R, Griffin AS, Manser M: **Selfish sentinels in cooperative mammals**. *Science* 1999, **284**(5420):1640-1644.
4. Dejonghe W, Berteloot E, Goris J, Boon N, Crul K, Maertens S, Hofte M, De Vos P, Verstraete W, Top EM: **Synergistic degradation of linuron by a bacterial consortium and isolation of a single linuron-degrading variovorax strain**. *Applied and environmental microbiology* 2003, **69**(3):1532-1541.
5. Axelrod R, Axelrod DE, Pienta KJ: **Evolution of cooperation among tumor cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(36):13474-13479.
6. Clutton-Brock T: **Cooperation between non-kin in animal societies**. *Nature* 2009, **462**(7269):51-57.
7. West SA, Griffin AS, Gardner A, Diggle SP: **Social evolution theory for microorganisms**. *Nat Rev Microbiol* 2006, **4**(8):597-607.
8. Bassler BL, Losick R: **Bacterially speaking**. *Cell* 2006, **125**(2):237-246.
9. Kassen R, Rainey PB: **The ecology and genetics of microbial diversity**. *Annual review of microbiology* 2004, **58**:207-231.
10. Rainey PB, Buckling A, Kassen R, Travisano M: **The emergence and maintenance of diversity: insights from experimental bacterial populations**. *Trends in ecology & evolution (Personal edition)* 2000, **15**(6):243-247.
11. Akiyama E, Kaneko K: **Dynamical systems game theory II. A new approach to the problem of the social dilemma**. *Physica D* 2002, **167**:36-71.
12. Brockhurst MA, Buckling A, Racey D, Gardner A: **Resource supply and the evolution of public-goods cooperation in bacteria**. *BMC biology* 2008, **6**:20.
13. Nowak MA: **Five rules for the evolution of cooperation**. *Science* 2006, **314**(5805):1560-1563.
14. Lenski R, Rose M, Simpson S, Tadler S: **Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations**. *The American naturalist* 1991, **138**(6):1315-1341.
15. Smith E, Morowitz HJ: **Universality in intermediary metabolism**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(36):13168-13173.
16. Ouzounis C, Kyrpides N: **The emergence of major cellular processes in evolution**. *FEBS letters* 1996, **390**(2):119-123.
17. Palsson BГ: **Systems Biology - Properties of Reconstructed Networks**. In.: Cambridge University Press; 2006.
18. Hardin G: **The competitive exclusion principle**. *Science* 1960, **131**:1292-1297.
19. Rosenzweig RF, Sharp RR, Treves DS, Adams J: **Microbial evolution in a simple unstructured environment: genetic differentiation in Escherichia coli**. *Genetics* 1994, **137**(4):903-917.
20. Treves DS, Manning S, Adams J: **Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli**. *Molecular biology and evolution* 1998, **15**(7):789-797.
21. Rozen DE, Lenski RE: **Long-Term Experimental Evolution in Escherichia coli. VIII. Dynamics of a Balanced Polymorphism**. *The American naturalist* 2000, **155**(1):24-35.
22. Turner EP, Souza V, Richard LE: **Tests of Ecological Mechanisms Promoting the Stable Coexistence of Two Bacterial Genotypes**. *Ecology* 1996, **77**(7):2119-2129.

23. Kurlandzka A, Rosenzweig RF, Adams J: **Identification of adaptive changes in an evolving population of Escherichia coli: the role of changes with regulatory and highly pleiotropic effects**. *Molecular biology and evolution* 1991, **8**(3):261-281.
24. Lenski RE: **Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium Escherichia coli**. *Plant Breeding Reviews* 2004, **24**:225-265.
25. Cooper TF, Rozen DE, Lenski RE: **Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(3):1072-1077.
26. Pelosi L, Kuhn L, Guetta D, Garin J, Geiselmann J, Lenski RE, Schneider D: **Parallel changes in global protein profiles during long-term experimental evolution in Escherichia coli**. *Genetics* 2006, **173**(4):1851-1869.
27. Cooper VS, Lenski RE: **The population genetics of ecological specialization in evolving Escherichia coli populations**. *Nature* 2000, **407**(6805):736-739.
28. Elena FS, Lenski RE: **Long-Term Experimental Evolution in Escherichia coli. VII. Mechanisms Maintaining Genetic Variability Within Population**. *Evolution; international journal of organic evolution* 1997, **51**(4):1058-1067.
29. Denamur E, Matic I: **Evolution of mutation rates in bacteria**. *Molecular microbiology* 2006, **60**(4):820-827.
30. Ochman H, Moran NA: **Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis**. *Science* 2001, **292**(5519):1096-1099.
31. Kærn Mads, Elston Timothy, Blake William, James C: **Stochasticity in gene expression: from theries to Phenotypes**. *Nature Reviews* 2005, **6**.
32. McAdams HH, Arkin A: **Stochastic mechanisms in gene expression**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(3):814-819.
33. Freilich S, Kreimer A, Borenstein E, Gophna U, Sharan R, Ruppin E: **Decoupling Environment-Dependent and Independent Genetic Robustness across Bacterial Species**. *PLoS Comput Biol*, **6**(2):e1000690.
34. Wang Z, Zhang J: **Abundant indispensable redundancies in cellular metabolic networks**. *Genome biology and evolution* 2009, **2009**:23-33.
35. Mahadevan R, Lovley DR: **The degree of redundancy in metabolic genes is linked to mode of metabolism**. *Biophysical journal* 2008, **94**(4):1216-1220.
36. Almaas E, Oltvai ZN, Barabasi AL: **The activity reaction core and plasticity of metabolic networks**. *PLoS Comput Biol* 2005, **1**(7):e68.
37. Kim PJ, Lee DY, Kim TY, Lee KH, Jeong H, Lee SY, Park S: **Metabolite essentiality elucidates robustness of Escherichia coli metabolism**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(34):13638-13642.
38. Matias Rodrigues JF, Wagner A: **Evolutionary plasticity and innovations in complex metabolic reaction networks**. *PLoS Comput Biol* 2009, **5**(12):e1000613.
39. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS *et al*: **Experimental determination and system level analysis of essential genes in Escherichia coli MG1655**. *Journal of bacteriology* 2003, **185**(19):5673-5684.
40. Joyce AR, Palsson BO: **Toward whole cell modeling and simulation: comprehensive functional genomics through the constraint-based approach**. *Progress in drug research Fortschritte der Arzneimittelforschung* 2007, **64**:265, 267-309.
41. Kauffman KJ, Prakash P, Edwards JS: **Advances in flux balance analysis**. *Current opinion in biotechnology* 2003, **14**(5):491-496.
42. Price ND, Reed JL, Palsson BO: **Genome-scale models of microbial cells: evaluating the consequences of constraints**. *Nat Rev Microbiol* 2004, **2**(11):886-897.
43. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions**. *Molecular systems biology* 2009, **5**:320.

44. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms**. *Nat Rev Microbiol* 2009, **7**(2):129-143.
45. Lee JM, Gianchandani EP, Papin JA: **Flux balance analysis in the era of metabolomics**. *Briefings in bioinformatics* 2006, **7**(2):140-150.
46. Mahadevan R, Edwards JS, Doyle FJ, 3rd: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophysical journal* 2002, **83**(3):1331-1340.
47. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Applied and environmental microbiology* 1994, **60**(10):3724-3731.
48. Borodina I, Krabben P, Nielsen J: **Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism**. *Genome research* 2005, **15**(6):820-829.
49. Edwards R, Glass L: **Combinatorial explosion in model gene networks**. *Chaos* 2000, **10**(3):691-704.
50. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network**. *Genome research* 2003, **13**(2):244-253.
51. Raman K, Chandra N: **Flux balance analysis of biological systems: applications and challenges**. *Briefings in bioinformatics* 2009, **10**(4):435-449.
52. Blazeck J, Alper H: **Systems metabolic engineering: Genome-scale models and beyond**. *Biotechnology journal*.
53. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli**. *Nature biotechnology* 2008, **26**(6):659-667.
54. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL: **Global organization of metabolic fluxes in the bacterium Escherichia coli**. *Nature* 2004, **427**(6977):839-843.
55. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli**. *Molecular systems biology* 2007, **3**:119.
56. Snitkin ES, Segre D: **Optimality criteria for the prediction of metabolic fluxes in yeast mutants**. *Genome informatics* 2008, **20**:123-134.
57. Edwards JS, Ibarra RU, Palsson BO: **In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data**. *Nature biotechnology* 2001, **19**(2):125-130.
58. Patil Kiran, Rocha Isabel, Förster Jochen, Jens N: **Evolutionary programming as a platform for in silico metabolic engineering**. *BMC bioinformatics [electronic resource]* 2005.
59. Rocha M, Maia P, Mendes R, Pinto JP, Ferreira EC, Nielsen J, Patil KR, Rocha I: **Natural computation meta-heuristics for the in silico optimization of microbial strains**. *BMC bioinformatics [electronic resource]* 2008, **9**:499.
60. Stephanopoulos G, Alper H, Moxley J: **Exploiting biological complexity for strain improvement through systems biology**. *Nature biotechnology* 2004, **22**(10):1261-1267.
61. Trinh CT, Carlson R, Wlaschin A, Srienc F: **Design, construction and performance of the most efficient biomass producing E. coli bacterium**. *Metabolic engineering* 2006, **8**(6):628-638.
62. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA: **Metabolic modeling of a mutualistic microbial community**. *Molecular systems biology* 2007, **3**:92.
63. Davison BH, Stephanopoulos G: **Coexistence of S. cerevisiae and E. coli in chemostat under substrate competition and product inhibition**. *Biotechnol Bioeng* 1986, **28**(11):1742-1752.
64. Hesseler J, Schmidt JK, Reichl U, Flockerzi D: **Coexistence in the chemostat as a result of metabolic by-products**. *Journal of mathematical biology* 2006, **53**(4):556-584.

65. Shou W, Ram S, Vilar JM: **Synthetic cooperation in engineered yeast populations**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(6):1877-1882.
66. Aledo JC, Perez-Claros JA, Esteban del Valle A: **Switching between cooperation and competition in the use of extracellular glucose**. *Journal of molecular evolution* 2007, **65**(3):328-339.
67. Hauert C, Michor F, Nowak MA, Doebeli M: **Synergy and discounting of cooperation in social dilemmas**. *Journal of theoretical biology* 2006, **239**(2):195-202.
68. Porcher E, Tenaillon O, Godelle B: **From metabolism to polymorphism in bacterial populations: a theoretical study**. *Evolution; international journal of organic evolution* 2001, **55**(11):2181-2193.
69. Doebeli M: **A model for the evolutionary dynamics of cross-feeding polymorphisms in microorganisms** *Population Ecology* 2002, **44**(2):59-70.
70. Strogatz SH: **Exploring complex networks**. *Nature* 2001, **410**(6825):268-276.
71. Sporns O, Chialvo DR, Kaiser M, Hilgetag CC: **Organization, development and function of complex brain networks**. *Trends in cognitive sciences* 2004, **8**(9):418-425.
72. Ulrik B, Sabine C: **Phylogenetic graph models beyond trees**. *Discrete Appl Math* 2009, **157**(10):2361-2369.
73. Szabσ G, Fαth G: **Evolutionary games on graphs**. *Physics Reports* 2007, **446**(4-6):97-216.
74. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.
75. Poncela J, Gomez-Gardenes J, Floria LM, Sanchez A, Moreno Y: **Complex cooperative networks from evolutionary preferential attachment**. *PLoS ONE* 2008, **3**(6):e2449.
76. Dorogovtsev SN, Mendes JFF: **Evolution of networks**. *Advances in Physics* 2002, **51**(4):1079-1187.
77. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nature biotechnology* 2006, **24**(4):427-433.
78. Tsiaras LV: **Algorithms for the analysis and visualization of biomedical networks**. *PhD Thesis* 2009(Computer science department, University of Crete, Greece).
79. Pavlopoulos GA, Wegener AL, Schneider R: **A survey of visualization tools for biological network analysis**. *BioData mining* 2008, **1**:12.
80. Cormen H, Thomas, Leiserson E, Charles, Rivest L, Ronald, Stein C: **Introduction to Algorithms**. Cambridge Massachusetts: MIT Press; 1990.
81. Gibbons A: **Algorithmic graph theory**. Cambridge: Cambridge University Press; 1985.
82. Barrat A, Barthélemy M, Vespignani A: **Dynamical Processes on Complex Networks**. Cambridge: Cambridge University Press; 2008.
83. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.
84. http://www.genome.jp/kegg/pathway.html: **Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database**.
85. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A: **The architecture of complex weighted networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(11):3747-3752.
86. Newman ME: **Analysis of weighted networks**. *Physical review* 2004, **70**(5 Pt 2):056131.
87. Freeman L: **Centrality in social networks: Conceptual clarification**. *Social Networks* 1979, **1**(3):215-239.
88. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**(6804):651-654.

89. Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T: **CFinder: locating cliques and overlapping modules in biological networks**. *Bioinformatics (Oxford, England)* 2006, **22**(8):1021-1023.
90. Yanai I, DeLisi C: **The society of genes: networks of functional links between genes from comparative genomics**. *Genome Biol* 2002, **3**(11):research0064.
91. Chiba N, Nishizeki T: **Arboricity and subgraph listing algorithms**. *SIAM Journal on Computing* 1985, **14**(1):210-223.
92. Ostergard RJP: **A New Algorithm for the Maximum-Weight Clique Problem**. *Electronic Notes in Discrete Mathematics, 6th Twente Workshop on Graphs and Combinatorial Optimization* 1999, **3**:153-156.
93. Fürer M: **A fatser algorithm for finding maximum independent sets in sparse graphs**. *LECTURE NOTES IN COMPUTER SCIENCE* 2006, **3887**:491-501.
94. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks**. *Nature* 1998, **393**(6684):440-442.
95. Kalna G, Higham JD: **A clustering coefficient for weighted networks, with application to gene expression data**. *AI Communications* 2007, **20**:263-271.
96. Newman ME: **Assortativity mixing in networks**. *Phys Rev Lett* 2002, **89**(20):208701.
97. Leung CC, Chau HF: **Weighted assortative and disassortative networks model**. *Physica A: Statistical Mechanics and its Applications* 2007, **378**(2):591-602.
98. Ravasz E: **Detecting hierarchical modularity in biological networks**. *Methods in molecular biology (Clifton, NJ* 2009, **541**:145-160.
99. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**(5586):1551-1555.
100. Erdős P, Rényi A: **On Random Graphs**. *Publicationes Mathematicae* 1959, **6**:290–297.
101. Garlaschelli D: **The weighted random graph model**. *New Journal of Physics* 2009, **11**.
102. Barabasi AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509-512.
103. Caldarelli G: **Scale-free networks**. Oxford: Oxford University press; 2007.
104. Caldarelli G, Capocci A, De Los Rios P, Munoz MA: **Scale-free networks from varying vertex intrinsic fitness**. *Physical review letters* 2002, **89**(25):258702.
105. Kumar R, Raghavan P, Rajagopalan S, Tomkins A: **Extracting large-scale knowledge bases from the web**. In: *VLDB'99: Proceedings of the 25th International Conference on Very Large Data Bases: 1999; San Francisco*: Morgan Kaufmann Publishers Inc.; 1999: 639-650.
106. Dangalchev C: **Generation models for scale-free networks**. *Physica A: Statistical Mechanics and its Applications* 2004, **338**(3-4):659-671.
107. Kalisky T, Sreenivasan S, Braunstein LA, Buldyrev SV, Havlin S, Stanley HE: **Scale-free networks emerging from weighted random graphs**. *Physical review* 2006, **73**(2 Pt 2):025103.
108. Servedio VD, Caldarelli G, Butta P: **Vertex intrinsic fitness: how to produce arbitrary scale-free networks**. *Physical review* 2004, **70**(5 Pt 2):056126.
109. Ravasz E, Barabasi AL: **Hierarchical organization in complex networks**. *Physical review* 2003, **67**(2 Pt 2):026112.
110. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761 Suppl):C47-52.
111. Joyce Andrew, Bernhard P: **The model organism as a system: integrating omics data sets**. *Nature* 2006, **7**.
112. Joshi A, Palsson BO: **Metabolic dynamics in the human red cell. Part I--A comprehensive kinetic model**. *Journal of theoretical biology* 1989, **141**(4):515-528.
113. Nakayama Y, Kinoshita A, Tomita M: **Dynamic simulation of red blood cell metabolism and its application to the analysis of a pathological condition**. *Theoretical biology & medical modelling* 2005, **2**(1):18.

114. Gershenson C: **Introduction to Random Boolean Networks**. *Workshop and Tutorial Proceedings, Ninth International Conference on the Simulation and Synthesis of Living Systems* 2004:160-173.
115. Kauffman SA: **Metabolic stability and epigenesis in randomly constructed genetic nets**. *Journal of theoretical biology* 1969, **22**(3):437-467.
116. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**(3-4):601-620.
117. Murphy K MS: **Modelling gene expression data using dynamic Bayesian networks**. In.: Computer Science Division, University of California, Berkeley; 1999.
118. David H: **A tutorial on learning with Bayesian networks**. In: *Learning in graphical models.* MIT Press; 1999: 301-354.
119. Akutsu T, Miyano S, Kuhara S: **Inferring qualitative relations in genetic networks and metabolic pathways**. *Bioinformatics (Oxford, England)* 2000, **16**(8):727-734.
120. Liang S, Fuhrman S, Somogyi R: **Reveal, a general reverse engineering algorithm for inference of genetic network architectures**. *Pac Symp Biocomput* 1998:18-29.
121. Nam DS, Seunghyun; Kim, Sangsoo: **An efficient top-down search algorithm for learning Boolean networks of gene expression**. *Machine Learning* 2006, **65**(1):229-245.
122. Lähdesmäki H: **On Learning Gene Regulatory Networks under the Boolean Network Model**. *Machine Learning* 2002, **52**:147-163.
123. Osamu Hirose NN, Yoshinori Tamada, Hideo Bannai, Seiya Imoto and Satoru Miyano: **Estimating Gene Networks from Expression Data and Binding Location Data via Boolean Networks**. In: *Computational Science and Its Applications – ICCSA 2005.* vol. 3482/2005: Springer Berlin / Heidelberg; 2005: 349-356.
124. Martin S, Zhang Z, Martino A, Faulon JL: **Boolean Dynamics of Genetic Regulatory Networks Inferred from Microarray Time Series Data**. *Bioinformatics (Oxford, England)* 2007.
125. Datta A, Choudhary A, Bittner ML, Dougherty ER: **External control in Markovian genetic regulatory networks: the imperfect information case**. *Bioinformatics (Oxford, England)* 2004, **20**(6):924-930.
126. Pal R, Datta A, Bittner ML, Dougherty ER: **Intervention in context-sensitive probabilistic Boolean networks**. *Bioinformatics (Oxford, England)* 2005, **21**(7):1211-1218.
127. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL: **A Bayesian approach to reconstructing genetic regulatory networks with hidden factors**. *Bioinformatics (Oxford, England)* 2005, **21**(3):349-356.
128. Bernard A, Hartemink AJ: **Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data**. *Pac Symp Biocomput* 2005:459-470.
129. Dojer N, Gambin A, Mizera A, Wilczynski B, Tiuryn J: **Applying dynamic Bayesian networks to perturbed gene expression data**. *BMC bioinformatics [electronic resource]* 2006, **7**:249.
130. Zou M, Conzen SD: **A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data**. *Bioinformatics (Oxford, England)* 2005, **21**(1):71-79.
131. Chen T, He HL, Church GM: **Modeling gene expression with differential equations**. *Pac Symp Biocomput* 1999:29-40.
132. de Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, Miyano S: **Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations**. *Pac Symp Biocomput* 2003:17-28.
133. D'Haeseleer P, Wen X, Fuhrman S, Somogyi R: **Linear modeling of mRNA expression levels during CNS development and injury**. *Pac Symp Biocomput* 1999:41-52.
134. Gustafsson M, Hornquist M, Lombardi A: **Constructing and analyzing a large-scale gene-to-gene regulatory network--lasso-constrained inference and**

biological validation. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 2005, **2**(3):254-261.
135. van Someren EP, Wessels LF, Reinders MJ: **Linear modeling of genetic networks from experimental data**. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB* 2000, **8**:355-366.
136. Wu FX, Zhang WJ, Kusalik AJ: **Modeling gene expression from microarray expression data with state-space equations**. *Pac Symp Biocomput* 2004:581-592.
137. Drulhe SF-T, G. de Jong, H. Viari, A.: **Reconstruction of Switching Thresholds in Piecewise-Affine Models of Genetic Regulatory Networks**. *LECTURE NOTES IN COMPUTER SCIENCE* 2006(3927):184-199.
138. Glass L, Kauffman SA: **The logical analysis of continuous, non-linear biochemical control networks**. *Journal of theoretical biology* 1973, **39**(1):103-129.
139. Radde N, Gebert J, Forst CV: **Systematic component selection for gene-network refinement**. *Bioinformatics (Oxford, England)* 2006, **22**(21):2674-2680.
140. Sorribas A, Curto R, Cascante M: **Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: model validation and dynamic behavior**. *Mathematical biosciences* 1995, **130**(1):71-84.
141. Vohradsky J: **Neural network model of gene expression**. *Faseb J* 2001, **15**(3):846-854.
142. Voit EO, Radivoyevitch T: **Biochemical systems analysis of genome-wide expression data**. *Bioinformatics (Oxford, England)* 2000, **16**(11):1023-1037.
143. Weaver DC, Workman CT, Stormo GD: **Modeling regulatory networks with weight matrices**. *Pac Symp Biocomput* 1999:112-123.
144. Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell**. *Science* 2002, **297**(5584):1183-1186.
145. Hasty J, Pradines J, Dolnik M, Collins JJ: **Noise-based switches and amplifiers for gene expression**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(5):2075-2080.
146. Spudich JL, Koshland DE, Jr.: **Non-genetic individuality: chance in the single cell**. *Nature* 1976, **262**(5568):467-471.
147. Gillespie DT: **Stochastic Simulation of Chemical Kinetics**. *Annu Rev Phys Chem* 2006.
148. Aranda JS, Salgado E, Munoz-Diosdado A: **Multifractality in intracellular enzymatic reactions**. *Journal of theoretical biology* 2006, **240**(2):209-217.
149. Schnell S, Turner TE: **Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws**. *Progress in biophysics and molecular biology* 2004, **85**(2-3):235-260.
150. Weiss M, Elsner M, Kartberg F, Nilsson T: **Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells**. *Biophysical journal* 2004, **87**(5):3518-3524.
151. Haseltine EL, Rawlings JB: **On the origins of approximations for stochastic chemical kinetics**. *The Journal of chemical physics* 2005, **123**(16):164115.
152. Salis H, Kaznessis Y: **Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions**. *The Journal of chemical physics* 2005, **122**(5):54103.
153. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U: **COPASI--a COmplex PAthway SImulator**. *Bioinformatics (Oxford, England)* 2006, **22**(24):3067-3074.
154. Snoep JL, Bruggeman F, Olivier BG, Westerhoff HV: **Towards building the silicon cell: a modular approach**. *Bio Systems* 2006, **83**(2-3):207-216.
155. Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S: **Integrative model of the response of yeast to osmotic shock**. *Nature biotechnology* 2005, **23**(8):975-982.

156. Schellenberger J, Park JO, Conrad TM, Palsson BO: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC bioinformatics [electronic resource]* 2010, **11**:213.
157. Varma A, Palsson BO: **Metabolic Capabilities of Escherichia coli: I. Synthesis of Biosynthetic Precursors and Cofactors**. *Journal of theoretical biology* 1993, **165**(4):477-502.
158. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic acids research* 1999, **27**(1):29-34.
159. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic acids research* 2006, **34**(Database issue):D354-357.
160. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A: **ExPASy: The proteomics server for in-depth protein knowledge and analysis**. *Nucleic acids research* 2003, **31**(13):3784-3788.
161. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes**. *Nucleic acids research* 2005, **33**(19):6083-6089.
162. Edwards JS, Palsson BO: **Systems properties of the Haemophilus influenzae Rd metabolic genotype**. *The Journal of biological chemistry* 1999, **274**(25):17410-17416.
163. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular systems biology* 2007, **3**:121.
164. Majewski RA, Domach MM: **Simple constrained-optimization view of acetate overflow in E. coli**. *Biotechnol Bioeng* 1990, **35**(7):732-738.
165. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model**. *Genome research* 2004, **14**(7):1298-1309.
166. Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, Baart G, Schwartz JM: **Integration of metabolic databases for the reconstruction of genome-scale metabolic networks**. *BMC systems biology* 2010, **4**:114.
167. Covert MW, Xiao N, Chen TJ, Karr JR: **Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli**. *Bioinformatics (Oxford, England)* 2008, **24**(18):2044-2050.
168. Lee JM, Gianchandani EP, Eddy JA, Papin JA: **Dynamic analysis of integrated signaling, metabolic, and regulatory networks**. *PLoS Comput Biol* 2008, **4**(5):e1000086.
169. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M: **Dynamic modeling of the central carbon metabolism of Escherichia coli**. *Biotechnol Bioeng* 2002, **79**(1):53-73.
170. Costa RS, Machado D, Rocha I, Ferreira EC: **Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations**. *Bio Systems*, **100**(2):150-157.
171. Usuda Y, Nishio Y, Iwatani S, Van Dien SJ, Imaizumi A, Shimbo K, Kageyama N, Iwahata D, Miyano H, Matsui K: **Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production**. *Journal of biotechnology*, **147**(1):17-30.
172. Covert Markus, Schilling Christophe, Bernhard P: **Regulation of Gene Expression in Flux Balance Models of Metabolism**. 2001:73-78.
173. Cakir T, Kirdar B, Ulgen KO: **Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks**. *Biotechnol Bioeng* 2004, **86**(3):251-260.
174. Klamt S, Stelling J: **Combinatorial complexity of pathway analysis in metabolic networks**. *Molecular biology reports* 2002, **29**(1-2):233-236.

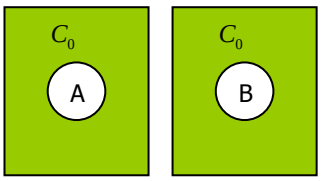
175. Carlson R, Fell D, Srienc F: **Metabolic pathway analysis of a recombinant yeast for rational strain development**. *Biotechnol Bioeng* 2002, **79**(2):121-134.
176. Wiback SJ, Palsson BO: **Extreme pathway analysis of human red blood cell metabolism**. *Biophysical journal* 2002, **83**(2):808-818.
177. Papin JA, Price ND, Edwards JS, Palsson BB: **The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy**. *Journal of theoretical biology* 2002, **215**(1):67-82.
178. Schilling CH, Edwards JS, Letscher D, Palsson BO: **Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems**. *Biotechnol Bioeng* 2000, **71**(4):286-306.
179. Henry CS, Broadbelt LJ, Hatzimanikatis V: **Thermodynamics-based metabolic flux analysis**. *Biophysical journal* 2007, **92**(5):1792-1805.
180. Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(21):7695-7700.
181. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN: **Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(31):12663-12668.
182. Herrgard MJ, Fong SS, Palsson BO: **Identification of genome-scale metabolic network models using experimentally measured flux profiles**. *PLoS Comput Biol* 2006, **2**(7):e72.
183. Herrgard MJ, Lee BS, Portnoy V, Palsson BO: **Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae**. *Genome research* 2006, **16**(5):627-635.
184. Feist AM, Scholten JC, Palsson BO, Brockman FJ, Ideker T: **Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri**. *Molecular systems biology* 2006, **2**:2006 0004.
185. Becker SA, Palsson BO: **Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation**. *BMC microbiology* 2005, **5**(1):8.
186. Almaas E, Oltvai Z, Barabasi A: **The Activity Reaction Core and Plasticity of Metabolic Networks**. *PloS Computational Biology* 2005, **1**(7).
187. Treves D, Manning S, Adams J: **Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli**. *Mol Biol Evol* 1998, **15**(7):789-797.
188. Ibarra RU, Edwards JS, Palsson BO: **Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth**. *Nature* 2002, **420**(6912):186-189.
189. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(23):15112-15117.
190. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models**. *Metabolic engineering* 2003, **5**(4):264-276.
191. Patil KR, Rocha I, Forster J, Nielsen J: **Evolutionary programming as a platform for in silico metabolic engineering**. *BMC bioinformatics [electronic resource]* 2005, **6**:308.
192. Murabito E, Simeonidis E, Smallbone K, Swinton J: **Capturing the essence of a metabolic network: a flux balance analysis approach**. *Journal of theoretical biology* 2009, **260**(3):445-452.
193. Lee S, Phalakornkule C, Domach MM, Grossmann IE: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**(2-7):711-716.

194. Holzhutter HG: **The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks**. *European journal of biochemistry / FEBS* 2004, **271**(14):2905-2922.
195. van Riel NA, Giuseppin ML, Verrips CT: **Dynamic optimal control of homeostasis: an integrative system approach for modeling of the central nitrogen metabolism in Saccharomyces cerevisiae**. *Metabolic engineering* 2000, **2**(1):49-68.
196. Palsson BO, Joshi A: **On the dynamic order of structured Escherichia coli growth models**. *Biotechnol Bioeng* 1987, **29**(6):789-792.
197. Barrett CL, Palsson BO: **Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach**. *PLoS Comput Biol* 2006, **2**(5):e52.
198. Herrgard MJ, Covert MW, Palsson BO: **Reconciling gene expression data with known genome-scale regulatory network structures**. *Genome research* 2003, **13**(11):2423-2434.
199. Akesson M, Forster J, Nielsen J: **Integration of gene expression data into genome-scale metabolic models**. *Metabolic engineering* 2004, **6**(4):285-293.
200. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO: **Integrating high-throughput and computational data elucidates bacterial networks**. *Nature* 2004, **429**(6987):92-96.
201. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)**. *Genome Biol* 2003, **4**(9):R54.
202. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox**. *Nature protocols* 2007, **2**(3):727-738.
203. http://www.gnu.org/software/glpk/: **GNU Linear Programming Kit**.
204. Qian ZG, Xia XX, Lee SY: **Metabolic engineering of Escherichia coli for the production of putrescine: a four carbon diamine**. *Biotechnol Bioeng* 2009, **104**(4):651-662.

Appendix

A. Here it is proved that in a spatially homogeneous environment consisting of a single limited resource, mass conservation does not allow relative and consequently absolute group benefit to emerge from two different populations that compete for the primal source alone independently of their initial frequency or the amount of the resource. This suggests that other sources of heterogeneity such as by-production might play a critical role in growth efficiency. The proof is given in two steps.

$$yield = \frac{\mu}{u}$$



1. Homogeneous growth dynamics.

Let A and B be two different cell populations. Let first the two populations grow independently on a limited resource of initial concentration C_0 and initial biomass concentration equal to b_0.

Let μ_A and u_A be the growth rate and the uptake rate respectively of the population A whereas μ_B and u_B be the growth rate and the uptake rate of the population B. Let T_A and T_B be the period of metabolism of the main source for population A and B respectively.

Assuming that the growth and uptake rates remain constant throughout this time period, the biomass concentration B_A and B_B just before the main resource becomes exhausted is given for each population by:

$B_A = b_0 \cdot e^{\mu_A T_A}$ and $B_B = b_0 \cdot e^{\mu_B T_B}$

On the other hand, the conservation of mass gives that:

$$C_0 = \frac{u_A}{\mu_A} b_0 (1 - e^{\mu_A T_A}) = \frac{u_B}{\mu_B} b_0 (1 - e^{\mu_B T_B}) \quad \Rightarrow$$

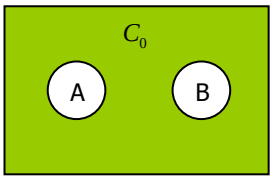
$$\frac{u_A}{\mu_A} \cdot \frac{\mu_B}{u_B} = \frac{(1 - e^{\mu_B T_B})}{(1 - e^{\mu_A T_A})} \qquad [A1]$$

For the population A to exhibit superior growth performance relative to B, B_A must be greater than B_B, which implies that:

$$B_A > B_B \quad \Rightarrow \quad b_0 \cdot e^{\mu_A T_A} > b_0 \cdot e^{\mu_B T_B} \quad \Rightarrow \quad 1 - e^{\mu_A T_A} < 1 - e^{\mu_B T_B} \quad \Rightarrow$$

$$\frac{1 - e^{\mu_B T_B}}{1 - e^{\mu_A T_A}} < 1 \quad \overset{[A1]}{\Rightarrow} \quad \frac{u_A}{\mu_A} \cdot \frac{\mu_B}{u_B} < 1 \qquad [A2]$$

In other words, the condition for a population A to perform better than a population B is: $yield_A > yield_B$



2. Heterogeneous growth dynamics.

Let the two populations A and B to co-grow on a limited resource of initial concentration C_0. Let $a \in [0,1]$ define their initial population ratio and let T_{AB} be the period of time until the main source is exhausted in the common medium.

Assuming that the growth and uptake rates remain constant throughout this time period, the group biomass concentration B_{AB} just before the main resource becomes exhausted is given by:

$$B_{AB} = a \cdot b_0 \cdot e^{\mu_A T_{AB}} + (1-a) \cdot b_0 \cdot e^{\mu_B T_{AB}}$$

Let population A exhibit superior growth performance relative to B in single-growth so that equation [A2] holds. To compare the growth performance of the group of A and B under competition with the performance of the best homogenous population (population A), the conservation of mass gives:

$$C_0 = \frac{u_A}{\mu_A} b_0 (1 - e^{\mu_A T_A}) = \frac{u_A}{\mu_A} a b_0 (1 - e^{\mu_A T_{AB}}) + \frac{u_B}{\mu_B} (1-a) b_0 (1 - e^{\mu_B T_{AB}}) \quad \Rightarrow$$

$$\lambda_A (1 - e^{\mu_A T_A}) = \lambda_A a (1 - e^{\mu_A T_{AB}}) + \lambda_B (1-a)(1 - e^{\mu_B T_{AB}}) \quad \Rightarrow$$

$$1 - e^{\mu_A T_A} = a (1 - e^{\mu_A T_{AB}}) + \frac{\lambda_B}{\lambda_A} (1-a)(1 - e^{\mu_B T_{AB}}) \quad \Rightarrow$$

$$e^{\mu_A T_A} = 1 - a (1 - e^{\mu_A T_{AB}}) - \frac{\lambda_B}{\lambda_A} (1-a)(1 - e^{\mu_B T_{AB}}) \quad [A3]$$

The terms λ_A and λ_B correspond to:

$$\lambda_A = \frac{u_A}{\mu_A} = \frac{1}{yield_A}$$

$$\lambda_B = \frac{u_B}{\mu_B} = \frac{1}{yield_B}$$

Based on [A2]: $\frac{\lambda_B}{\lambda_A} > 1$

For the group population of A and B to exhibit superior growth performance relative to the best homogeneous growth performance A, B_{AB} must be greater than B_A, which implies that:

$$B_{AB} > B_A \quad \Rightarrow \quad a \cdot b_0 \cdot e^{\mu_A T_{AB}} + (1-a) \cdot b_0 \cdot e^{\mu_B T_{AB}} > b_0 \cdot e^{\mu_A T_A} \quad \Rightarrow$$

$$a \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > e^{\mu_A T_A} \quad \overset{[A3]}{\Rightarrow}$$

$$a \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > 1 - a(1-e^{\mu_A T_{AB}}) - \frac{\lambda_B}{\lambda_A}(1-a)(1-e^{\mu_B T_{AB}}) \quad \Rightarrow$$

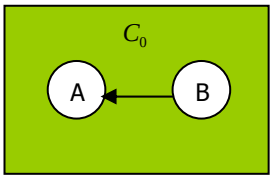
$$a \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > 1 - a + a \cdot e^{\mu_A T_{AB}} - \frac{\lambda_B}{\lambda_A} + \frac{\lambda_B}{\lambda_A} e^{\mu_B T_{AB}} + a\frac{\lambda_B}{\lambda_A} - a\frac{\lambda_B}{\lambda_A} e^{\mu_B T_{AB}} \quad \Rightarrow$$

$$(1-a) \cdot (1-\frac{\lambda_B}{\lambda_A}) \cdot e^{\mu_B T_{AB}} > (1-a) \cdot (1-\frac{\lambda_B}{\lambda_A}) \quad [\mathrm{A4}] \quad \overset{[A2]}{\Rightarrow}$$

$$e^{\mu_B T_{AB}} < 1 \quad [\mathrm{A5}]$$

The inequality [A5] is impossible implying that B_{AB} can never be greater than B_A. Competition alone cannot lead to group benefit.

On the other hand, the equality $B_{AB} = B_A$ occurs if and only if $\frac{\lambda_B}{\lambda_A} = 1$ or $a = 1$ as derived from [A4], which underlies that, either the two competing populations must exhibit the same growth yield or the trivial solution that the population consists of the population A and only.



B. A simple cross-feeding example is shown here. The cross-feeding interactions and the assumptions presented in the following are for the sake of simplicity of the dynamic equations. The aim is to provide an example among many scenarios that allow by-production utilization, where group benefit can emerge.

The population B is capable of by-producing a metabolite, which it cannot consume but which the population A can utilize. The conditions under which group benefit can emerge are described.

The homogenous population dynamics of each population do not change. Thus, for the population A to perform better than B equation [A2] must hold. However, the dynamics of the mixed population change. The metabolite s provided by the population B in the growth medium is produced as long as the main source is metabolized. After that period, which equals to T_{AB}, the utilization of s from population A occurs. The period of time until the s is exhausted in the common medium is denoted as T^s_{AB}. The growth rate of population A during T^s_{AB} is assumed to be constant and equal to μ^s_A. Sequential consumption of the resources is assumed for population A. Therefore, the equation [A3] holds.

The group biomass concentration B_{AB} is thus given by:

$$B_{AB} = a \cdot b_0 \cdot e^{\mu_A T_{AB}} e^{\mu_A^s T_{AB}^s} + (1-a) \cdot b_0 \cdot e^{\mu_B T_{AB}}$$

If population B is present ($a \neq 1$) and population A can utilize s then the new term $K = e^{\mu_A^s T_{AB}^s} > 1.$

For the group population of A and B to exhibit superior growth performance relative to the best homogeneous growth performance A, B_{AB} must be greater than B_A, which implies that:

$$B_{AB} > B_A \quad \Rightarrow \quad a \cdot b_0 \cdot e^{\mu_A T_{AB}} \cdot K + (1-a) \cdot b_0 \cdot e^{\mu_B T_{AB}} > b_0 \cdot e^{\mu_A T_A} \quad \Rightarrow$$

$$a \cdot K \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > e^{\mu_A T_A} \quad \overset{[A3]}{\Rightarrow}$$

$$a \cdot K \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > 1 - a(1 - e^{\mu_A T_{AB}}) - \frac{\lambda_B}{\lambda_A}(1-a)(1 - e^{\mu_B T_{AB}}) \quad \Rightarrow$$

$$a \cdot K \cdot e^{\mu_A T_{AB}} + (1-a) \cdot e^{\mu_B T_{AB}} > 1 - a + a \cdot e^{\mu_A T_{AB}} - \frac{\lambda_B}{\lambda_A} + \frac{\lambda_B}{\lambda_A} e^{\mu_B T_{AB}} + a\frac{\lambda_B}{\lambda_A} - a\frac{\lambda_B}{\lambda_A} e^{\mu_B T_{AB}} \quad \Rightarrow$$

$$a \cdot (K-1) \cdot e^{\mu_A T_{AB}} + (1-a) \cdot (1 - \frac{\lambda_B}{\lambda_A}) \cdot e^{\mu_B T_{AB}} > (1-a) \cdot (1 - \frac{\lambda_B}{\lambda_A}) \quad \Rightarrow$$

$$a \cdot (K-1) \cdot e^{\mu_A T_{AB}} > (1-a) \cdot (1 - \frac{\lambda_B}{\lambda_A}) \cdot (1 - e^{\mu_B T_{AB}}) \quad \Rightarrow$$

$$K > 1 + \frac{(1-a)}{a} \cdot (1 - \frac{\lambda_B}{\lambda_A}) \cdot \frac{(1 - e^{\mu_B T_{AB}})}{e^{\mu_A T_{AB}}} > 1 \qquad [A6]$$

The inequality [A6] defines the condition under which group benefit can emerge. The left-hand term K, which describes the benefit from cross-feeding and depends on the amount of s that the population B can provide and the growth rate of population A during the consumption of s, must exceed the right-hand term. The condition also indicates dependence on the initial population ratio (here, $a \in (0,1)$) and the growth yields of the two populations with respect to the primal source.

This work investigates whether conditions like [A6] can hold in competing populations comprised of genetically perturbed cells given genome-scale descriptions of the cells and a constraint-based approach to describe their dynamics. In simulations, the populations are not restricted to utilize the by-products sequentially or exchange by-products in one direction. The populations are allowed to dynamically adapt and utilize the given environment according to their metabolic capabilities.