ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ - ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΙΔΡΥΜΑ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΡΕΥΝΑΣ
ΙΝΣΤΙΤΟΥΤΟ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Προδιαγράφοντας την δομή μιας πρωτεΐνης στην αλληλουχία της

(αποτελέσματα που προκύπτουν για ένα απλό δομικό πρότυπο)

Υποβάλλων:
ΠΑΛΙΑΚΑΣΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ
Βιολόγος - Πτυχ. Τμ. Βιολ./Πανεπ. Αθηνών 1989
Κάτοχος Μεταπτυχ. Τίτλου Ειδίκευσης Τμ. Βιολ./Πανεπ. Κρήτης 1992
Υποψήφιος Διδάκτωρ Τμ. Βιολ./Πανεπ. Κρήτης

Επιβλέπων:
Μ.ΚΟΚΚΙΝΙΔΗΣ
Αν. Καθηγητής Τμήμ. Βιολογίας
Πανεπιστημίου Κρήτης

ΥΠΟΒΟΛΗ: ΜΑΡΤΙΟΣ 2001

UNIVERSITY OF CRETE
DEPARTMENT OF BIOLOGY

FOUNDATION OF RESEARCH AND TECHNOLOGY-HELLAS (FORTH)
INSTITUTE OF MOLECULAR BIOLOGY AND BIOTECHNOLOGY (IMBB)

Ph.D. Thesis

Specification of Protein Structure in Protein Sequence

(results for a simple tertiary structural motif)

By:
PALIAKASIS CONSTANTINE
Biologist - First Degree: Dept.Biology/Univ. of Athens 1989
M.Sc.: Dept.Biology/Univ. of Crete 1992
Thesis submitted to: Dept.Biology/Univ. of Crete

Supervisor:
M.KOKKINIDIS
Assoc. Prof./Dept. Biology
University of Crete

SUBMISSION: MARCH 2001

Συμβουλευτική Επιτροπή:	Κοκκινίδης Μ., Αν. Καθηγητής (Επιβλέπων)
	Μπουριώτης Β., Καθηγητής
	Ηλιόπουλος Η., Αν. Καθηγητής
Εξεταστική Επιτροπή:	Κοκκινίδης Μ., Αν. Καθηγητής (Επιβλέπων)
	Μπουριώτης Β., Καθηγητής
	Ηλιόπουλος Η., Αν. Καθηγητής
	Πανόπουλος Ν., Καθηγητής
	Παπαματθαιάκης Ι., Καθηγητής
	Αλεξανδράκη Δ., Αν. Καθηγήτρια
	Χαλεπάκης Γ., Αν. Καθηγητής

Πρόλογος

Το αντικείμενο της εργασίας, που περιγράφεται στο παρόν σύγγραμμα, κινείται στο χώρο των χαρακτηριστικών που πρέπει να έχει η αμινοξική αλληλουχία μιας πρωτεΐνης, ώστε να μπορεί να υπάρξει στην σωστή (λειτουργική) τρισδιάστατη δομή, θα πρέπει δε να είναι εύκολα κατανοητό για οποιονδήποτε έχει βασικές γνώσεις βιοχημείας και -ίσως- μοριακής εξέλιξης. Από την άλλη, όπως θα πρέπει να είναι προφανές και από τον τίτλο, η εργασία αυτή δεν ασχολείται με μηχανισμούς ή/και διαδρομές που οδηγούν από την κατάσταση αποδιάταξης στη φυσική δομή.

Η εργασία, στο ερευνητικό της μέρος, πραγματοποιήθηκε ολόκληρη στο εργαστήριο Κρυσταλλογραφίας Πρωτεϊνών του Τμήματος Βιολογίας της Σχολής Θετικών Επιστημών του Πανεπιστημίου Κρήτης, κατά τα ακαδημαϊκά έτη 1989-1995, με υποτροφία που παραχωρήθηκε ευγενικά από το Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας του Ιδρύματος Τεχνολογίας και Έρευνας. Στη μορφή της παρούσας διατριβής, γράφτηκε κατά ένα μέρος στο Ηράκλειο της Κρήτης, στα τέλη του 1995, και κατά ένα μέρος στην Αθήνα στα μέσα του 1998, κατόπιν υποχρεωτικής διακοπής εξαιτίας στράτευσης, για τα έτη 1996-1997. Υποβλήθηκε όμως στο Τμήμα Βιολογίας του Πανεπιστημίου Κρήτης, για την απόκτηση διδακτορικού διπλώματος, δύομισυ έτη αργότερα, τον Μάρτιο του 2001, αφ'ενός εξαιτίας κάποιων τυπικών κωλυμάτων, και αφ'ετέρου σαν συνέπεια του καταστροφικού σεισμού που έπληξε την Αθήνα στις 7 Σεπ 1999 και προβλημάτων υγείας.

Όπως διαπίστωσα από νωρίς και με μεγάλη μου χαρά, η σειρά εκείνη, που είναι η καλύτερη για την εννοιολογικά αρτιότερη παρουσίαση της εργασίας, συμπίπτει με την ιστορική σειρά των αντίστοιχων ερευνητικών βημάτων. Πράγματι, και περίπου ασυναίσθητα, η εργασία εξελίχθηκε από το απλούστερο και το γενικότερο αλλά ασαφέστερο, προς το πιο συγκεκριμένο που γινόταν πιο πλήρες αλλά και πιο πολύπλοκο. Σε ένα πιο προσωπικό επίπεδο, η περιγραφή με ιστορική σειρά δίνει την ευκαιρία μιας παράλληλης παρουσίασης εκείνου που αποτελεί την ουσία μιας προσωπικής πνευματικής διαδρομής, στο τέλος της οποίας -επιπλέον- υπήρξε μια απρόσμενη ευχάριστη έκπληξη. Έτσι, για να δικαιολογήσει το σύγγραμμα την ύπαρξή του -και τον κόπο για να το διαβάσει κάποιος- δεν αρκεί μόνο η τυπική υποχρέωσή του να δώσει τη βάση για παραπέρα προβληματισμό, αλλά θα πρέπει μέχρι το τέλος να έχει μεταφέρει και το συναίσθημα αυτής της ευχάριστης έκπληξης.

Το πρόβλημα, στην επίλυση του οποίου καλείται να συνεισφέρει η εξεταζόμενη εργασία, τίθεται πιστεύω επαρκώς και με σαφήνεια στη Γενική Εισαγωγή. Σε συντομία, ενώ γενικά η οργάνωση των πρωτεϊνικών δομών με στερεοδιαταξικούς όρους και σε διάφορα επίπεδα είναι απλή και κατανοητή, η πρόγνωση της τρισδιάστατης δομής κάθε συγκεκριμένης πρωτεϊνικής αλυσίδας από την αλληλουχία των αμινοξέων έχει αποδειχθεί εξαιρετικά δύσκολη. Το 1989-90 το καλύτερο που μπορούσε να κάνει κάποιος, δίχως κάποια ομόλογη, πειραματικά προσδιορισμένη δομή, κινούταν ανάμεσα σε ό,τι μπορεί να προσφέρει η στατιστική πρόγνωση της δευτεροταγούς δομής, και το διάγραμμα υδροφοβικότητας.

Δείχνεται, λοιπόν, για ποιο λόγο το κλειδί, για τη λύση του προβλήματος αυτού, βρίσκεται στο γεγονός ότι κάθε πρωτεϊνική δομή οργανώνεται με βάση ένα από λίγα (ίσως περίπου 100) δομικά πρότυπα. Τα πρότυπα αυτά, χωρίς να αναφέρονται στις λεπτομέρειες της κάθε δομής, περιγράφουν τρόπους με τους οποίους τα στοιχεία της δευτεροταγούς δομής (κυρίως α-έλικες και β-κλώνοι) συγκροτούνται σε "χρήσιμες" τριτοταγείς δομές. Το δομικό πρότυπο που επιλέχθηκε, το δεμάτι των 4 α-ελίκων, είναι από τα συχνά απαντώμενα και μάλιστα σε οικογένειες πρωτεϊνών ποικίλης λειτουργίας ή προέλευσης· ενώ, η απλότητά του και το γεγονός ότι συνήθως είναι ιδιαίτερα σταθερό, έχουν αποτελέσει τη βάση της συχνής και από ενωρίς εμφάνισής του σε πειράματα ανάλυσης και σχεδιασμού.

Τα Κεφάλαια Α.Ι και Α.ΙΙ περιγράφουν το είδος των αναλύσεων που έγιναν από τον Δεκέμβρη του 1989 εως τον Σεπτέμβρη του 1992, τα αποτελέσματα των οποίων έχουν σε μεγάλο βαθμό δημοσιευτεί. Κάθε θέση σε ένα δομικό πρότυπο, προβάλλει (κυρίως λόγω πακεταρίσματος με θέσεις γειτονικές στο χώρο) περιορισμούς, συχνά συμβατούς με λίγα μόνο από τα είκοσι αμινοξέα, ενώ ακόμη και τα συνδετικά τμήματα μεταξύ των α-ελίκων, που συνήθως περιγράφονται με τον ατυχή όρο "τυχαία δομή", παρουσιάζουν συγκεκριμένες προτιμήσεις. Το σύνολο των προτιμήσεων όλων των τμημάτων του δομικού προτύπου, όπως πινακοποιούνται στα κεφάλαια αυτά, αποτελεί τη βάση για τη συμβατότητά του με κάποια αλληλουχία και ανοίγει το δρόμο για την κατανόηση των αρχών που διέπουν την ταξινόμηση -στην αλληλουχία- των πληροφοριών που αφορούν τη δομή[1].

Το Μέρος Β περιέχει τις άμεσες εφαρμογές πινάκων, όπως αυτοί των κεφαλαίων του Μέρους Α (κυρίως σε μη ομόλογη κατασκευή μοντέλων και σχεδιασμό μεταλλαγών, εργασία που έγινε στα τέλη του 1992), καθώς και τις έμμεσες συνέπειες και γενικότερες επιπτώσεις τους σε θέματα δομής πρωτεϊνών και εξέλιξής τους στο χρόνο (όπως διερευνήθηκαν κατά τα έτη 1993-Ιούλιο 1995). Όσον αφορά τις άμεσες εφαρμογές, αρχικά περιγράφεται σε λεπτομέρεια μια διαδικασία πρωτεϊνικού σχεδιασμού (protein design), με ζητούμενο τη διαίρεση μιας δομικής ενότητας (structural domain) στα δύο, που όμως έμεινε χωρίς πειραματική επιβεβαίωση (αφού η υλοποίησή του το καλοκαίρι του 1994 διακόπηκε, από αιτίες εξωγενείς προς τον γράφοντα). Ως έχει, μαζί με κάποια επιπλέον πειράματα που θα μπορούσαν να προετοιμαστούν με τον ίδιο τρόπο (και δίνονται πιο συνοπτικά), δείχνει τον τρόπο με τον οποίο τέτοιες διαδικασίες θα μπορούσαν, αφ'ενός να διευκολυνθούν στην θεωρητική ανάλυση πριν την κατάστρωση του πειραματικού τους μέρους και αφ'ετέρου να έχουν μειωμένο κίνδυνο αποτυχίας, χάρη στην εκλογίκευση από χρήση πληροφορίας του είδους. Δείχνει ακόμη το πείραμα αυτό τα όρια των πινάκων αυτών, καθώς και τα επιπλέον στοιχεία που χρειάζονται ώστε να προκύπτει μια πλήρης και αυτοδύναμη διαδικασία. Πάντως, εργασίες σε παρόμοια πειράματα σχεδιασμού πρωτεϊνών, στη βάση τέτοιων πινάκων, που δημοσιεύτηκαν πολύ αργότερα από άλλες ερευνητικές ομάδες, δείχνουν την πληροφοριακή ισχύ των πινάκων αυτών, και αναφέρονται λεπτομερώς στο αντίστοιχο κεφάλαιο.

[1] Η αλληλουχία προφανώς φέρει κι'άλλες πληροφορίες, σχετικές με λειτουργία, αλληλεπίδραση με άλλα μόρια, ανοσολογική συμπεριφορά, ή που απλά αποτελούν κληρονομιά καθαρά εξελικτικής προελεύσεως.

Πάντα στο Μέρος Β, και όσον αφορά τις γενικότερες συνέπειες, η διευρεύνηση της στατιστικής σταθερότητας των ίδιων πινάκων, η προσπάθεια εκμετάλλευσής τους για πρόβλεψη, και η σύγκριση με (ή καλύτερα: οι διαφορές από) τα αποτελέσματα από παρόμοιες στατιστικές αναλύσεις σε συναφή πρότυπα (δημοσιευμένες από άλλες ομάδες, αργότερα ή το πολύ παράλληλα), αποκαλύπτει τη φύση τους (την "ουσία" τους). Έτσι, αν και αναλύεται ένα μόνο δομικό πρότυπο, δεν αποκλείεται να γενικεύονται τα περισσότερα από τα συμπεράσματα, που προκύπτουν μέσα από τις ιδέες που συζητούνται εκεί. Παραπροϊόν της διαδικασίας αυτής: ένα πρόγραμμα σύγκρισης αλληλουχιών.

Η παρουσίαση της εργασίας άπτεται θεμάτων, που συνήθως απαντώνται σε βιβλιογραφικά διαφορετικούς χώρους. Καθώς καθένας από αυτούς τους χώρους είναι τεράστιος, θα ήταν ανόητο να επιχειρηθεί η πλήρης βιβλιογραφική κάλυψη όλων των χώρων· άλλωστε, σκοπός ενός διδακτορικού δεν είναι η απλή ανασκόπηση των εργασιών άλλων ερευνητών και μελετητών. Έτσι δίνονται είτε σημαντικά "ιστορικά" άρθρα που έθεσαν τις βάσεις για κάτι, είτε (όπου κάποια πράγματα έχουν ανατραπεί ή είναι πιο πρόσφατα) τα πιο επίκαιρα. Εξάλλου, έχει καταβληθεί κάθε προσπάθεια ώστε, ακόμη και όταν δεν υπάρχουν σαφείς βιβλιογραφικές παραπομπές (πχ βασικές γνώσεις φυσικής), να είναι σαφές ποια δεδομένα και άλλες πληροφορίες δεν εξήχθησαν από την εργασία αυτή.

Καθώς οι περισσότερες από τις γνώσεις μου σε θέματα προγραμματισμού Η/Υ προέρχονται από βιβλία, κυρίως μορφής εγχειριδίου, σχεδιασμένα για φοιτητές πανεπιστημίων των ΗΠΑ, πολλά από τα στοιχεία οργάνωσης του παρόντος συγγράμματος έχουν μεταφερθεί από εκεί, με στόχο τη "χρηστικότερη" δυνατή μορφή. Για παράδειγμα, η αρίθμηση των εικόνων και των πινάκων γίνεται μέσα στο κάθε κεφάλαιο: η Εικ. Α.Ι.4 είναι η Εικόνα 4 του Κεφαλαίου Ι του Μέρους Α. Έτσι, γνωρίζοντας κάποιος τι πραγματεύεται το Κεφ. Α.Ι, μπορεί να γνωρίζει τι περίπου περιέχει η εικόνα, χωρίς καν να χρειαστεί να γυρίσει εκεί, κάτι που θα διέκοπτε τη ροή της ανάγνωσης. Ενώ σε μια ενιαία αρίθμηση, για να θυμηθεί κανείς τι δείχνει η Εικόνα 14, που όμως βρίσκεται σε κάποιο άλλο κεφάλαιο, θα έπρεπε να πάρει τις εικόνες με τη σειρά. Το ίδιο σχήμα έχει χρησιμοποιηθεί και για τις σελίδες, τόσο για λόγους ομοιομορφίας μιας ενιαίας παρουσίασης, όσο και για τη διευκόλυνση του αναγνώστη: αν αναζητώντας κάτι που είναι στο Κεφ. Β.ΙΙ βλέπει στην κορυφή της σελίδας την ένδειξη Β.Ι.13 γνωρίζει προς τα που πρέπει να κινηθεί· ενώ σε μια ενιαία αρίθμηση από σελίδα 1 μέχρι 100, πρέπει να αναζητά ενδείξεις στο κείμενο για το που βρίσκεται προκειμένου να αποφασίσει προς τα που πρέπει να κινηθεί. Υπό το πρίσμα και της αναγνωσιμότητας, τα μαθηματικά (και γενικά συζητήσεις, που απαιτούν πιο εξειδικευμένες γνώσεις εκτός του χώρου της Βιολογίας) έχουν περιοριστεί στο ελάχιστο, και έχουν αντικατασταθεί με απλά λογικά επιχειρήματα. Εξάλλου, παράγραφοι με ψιλά γράμματα, ανάμεσα σε δύο κατακόρυφες γραμμές, μπορούν σε μια γρήγορη ανάγνωση να παραλειφθούν, χωρίς να χάνεται το νόημα, καθώς αποτελούν τεχνικές -κυρίως- ή άλλες λεπτομέρειες.

Μια ακόμη λεπτομέρεια, που χρήζει σχολιασμού, πηγάζει από την εξάρτηση της εργασίας αυτής από τους ηλεκτρονικούς υπολογιστές. Τα ισχυρότερα επιτραπέζια συστήματα της εποχής που ξεκίνησε η εργασία (1990), βασίζονταν στον επεξεργαστή Intel 80386 με 2-4MB κεντρικής μνήμης, 40MB σκληρό δίσκο και τις προβληματικότατες εκδόσεις των Windows 3.1, ενώ οι επεξεργαστές Pentium/90MHz εμφανίστηκαν ενώ έφευγα από το Ηράκλειο. (Θυμάμαι με τι χαρά είχαμε υποδεχτεί στην ομάδα το 1992 τέσσερα Acer/486SX/20MHz με 2-4MB μνήμη και λιγότερο από 100MB σκληρό δίσκο - το SX σημαίνει ότι ήταν μια ελαφριά έκδοση του 486, χωρίς τον εσωτερικό μαθηματικό συνεπεξεργαστή 487, και όχι απόλυτα 32bit αρχιτεκτονική στο εσωτερικό του…) Τα εργαστήρια ήσαν εξοπλισμένα με MicroVax II της Digital Equipment Corporation (DEC) κάτω από VAX/VMS 4.x και σκληρούς δίσκους 850MB, με ταχύτητα επεξεργαστή συγκρίσιμη με ενός υπολογιστή IBM 386/25MHz, και χαμηλότερης πραγματικής, αφού ο χρόνος μοιραζόταν σε περισσότερους από έναν χρήστες· ενώ οι υπολογιστές Convex, σε όποια εργαστήρια υπήρχαν, αποτελούσαν επένδυση και ήταν η αιχμή του δόρατος. Τα πρώτα Silicon Graphics στην φτωχή μορφή του Indy ήρθαν στην ομάδα μόλις το 1995, ενώ μέχρι τότε στην ομάδα είχαν έρθει σταδιακά από το 1991 κάποια MicroVax 3000 και μικρά VaxStation με μνήμες από 4MB μέχρι 12MB. Οι καλύτερες οθόνες 21” είχαν αναλύσεις 1280x1024 (1024x768 για πολλά Indy) και τα γυαλιά στερεοσκοπικής όρασης ρυθμούς ανανέωσης 50Hz μικτό (25Hz πραγματικό!). Έτσι, απαιτητικοί υπολογισμοί (πχ. μοριακής δυναμικής) που σήμερα γίνονται ρουτινοειδώς σε επιτραπέζια συστήματα, τότε απαιτούσαν ειδική άδεια σε κεντρικά μηχανήματα κάποιου υπολογιστικού κέντρου. Όχι λιγότερο όμως, το κατά πόσον τα αποτελέσματα της παρούσας δεν ανατράπηκαν και δεν παλιώσαν είναι ένα ακόμη μέτρο της επιτυχούς διεξαγωγής της.

Ευχαριστίες

Αναφέρθηκε ήδη ότι η εργασία, που περιγράφεται στο παρόν σύγγραμμα, πραγματοποιήθηκε ολόκληρη στο εργαστήριο Κρυσταλλογραφίας Πρωτεϊνών, του Τμήματος Βιολογίας, της Σχολής Θετικών Επιστημών του Πανεπιστημίου Κρήτης (ΠΚ), με υποτροφία που παραχωρήθηκε ευγενικά από το Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας (IMBB), του Ιδρύματος Τεχνολογίας και Έρευνας (ΙΤΕ). Ο επικεφαλής του εργαστηρίου Κρυσταλλογραφίας Πρωτεϊνών, τότε Επίκουρος Καθηγητής (και αργότερα και μέχρι πρόσφατα Αναπληρωτής Καθηγητής) του ΠΚ, Κοκκινίδης Μιχαήλ, ήταν ο κύριος επιβλέπων, ιδιαίτερα στο κομμάτι που αφορά την ανάλυση του δομικού προτύπου, ενώ σημαντική ήταν η συμβολή του και στην εξέλιξη των κομματιών εκείνων, που ξέφευγαν κάπως από το άμεσο γνωστικό πεδίο του.

Ο ερευνητής του IMBB/ITE Κυριάκος Πετράτος αξίζει όλα τα ευχαριστώ για τις -συνήθως μακροσκελείς- συζητήσεις σε θέματα δομής και λειτουργίας πρωτεϊνών, όπως αυτά προκύπτουν από τις κρυσταλλογραφικές αναλύσεις, και στα οποία θέματα η εμμονή του στη λεπτομέρεια και την ακρίβεια προσέδιδαν πάντα ιδιαίτερη ομορφιά. Επιπλέον, οι συζητήσεις με τη Μεταξία Βλάσση, και τους Γιάννη Παπανικολάου και Αλέκο Αθανασιάδη ήταν -συχνά- αποφασιστικής σημασίας για την αποσαφήνιση εννοιών, που αλλιώς θα είχα (τουλάχιστο) πολύ άσχημα παρερμηνεύσει, ενώ σημαντική ήταν και η άμεση πρόσβαση σε διάφορα δεδομένα (ή άλλες πηγές) που μου παρείχαν. Ο Δημήτρης Παρασκευής, τόσο στα πλαίσια της μεταπτυχιακής του εργασίας, όσο και με συζητήσεις και ερωτήσεις που ξαφνιάζουν, αποτέλεσε τον κύριο μοχλό γενίκευσης των αποτελεσμάτων, όπου ήταν δυνατό. Τέλος, οι διαρκείς συνεννοήσεις με τον Μανώλη Πιτταροκοίλη και την Renate Gessmann βοήθησαν να αρθούν σε μεγάλο βαθμό οι περιορισμοί σε υπολογιστική ισχύ, που αναφέρθηκαν νωρίτερα. Το πείραμα σχεδιασμού, αν εκτελέστηκε μέχρι κάποιο σημείο, το οφείλει στους συναδέλφους Τάσο Γεωργακόπουλο και Νεκτάριο Ταβερναράκη, με σημαντική συμβολή από τον Γιάννη Παπανικολάου και την Ντίνα Κοτσυφάκη. Από τις ίδιες γραμμές, θέλω να ευχαριστήσω ιδιαίτερα τους παραπάνω συναδέλφους και φίλους, αλλά και όλους εκείνους που πέρασαν από το εργαστήριο αυτό, τη μακρά αυτή περίοδο, που έκαναν χαρακτηριστικό το εργαστηρίου αυτού μια ευχάριστη και πάντα ιδιαίτερη και γόνιμη ατμόσφαιρα.

Οι κοπέλλες της Γραμματείας του Τμήματος Βιολογίας και -ας μου επιτραπεί- ιδιαίτερα η κα Χαρά Ζαφειροπούλου-Σφακιανάκη, όπως και της Γραμματείας του IMBB -και ας μου επιτραπεί ξανά- ιδιαίτερα η κα Γεωργία Χουλάκη, αξίζουν "εύφημο μνεία" για την υπομονή τους μαζί μου, τόσο για τη χρονική διάρκεια που τις ταλαιπώρησα, όσο και για το είδος και την έκταση των θεμάτων που με βοήθησαν να χειριστώ αποτελεσματικά.

Για τους γονείς μου, ας μου επιτραπεί να εκφράσω την ευγνωμοσύνη που χωρά ένας ολόκληρος κόσμος, για την στήριξη που μου παρείχαν όλα αυτά τα χρόνια -συχνά κάτω από δύσκολες γι'αυτούς συνθήκες- και χωρίς την οποία ένα διδακτορικό για μένα θα ήταν τουλάχιστον ουτοπία.

Τελευταίους, αλλά όχι λιγότερο σημαντικούς, από τη θέση αυτή θα ήθελα να ευχαριστήσω εκείνους που -κάτω από συνθήκες που συνήθως είναι ιδιαίτερα αντίξοες- είχαν τα κότσια να δημιουργήσουν κάτι νέο, σε ένα περιβάλλον που -δυστυχώς- συνήθως δεν μας εκπλήσσει ευχάριστα. Η εργασία που περιγράφεται παρακάτω φιλοδοξεί να εντάσσεται στα πλαίσια της σωστής και καλώς εννοούμενης εκμετάλλευσης των καρπών *και* της δικής τους προσπάθειας.

Θα ήθελα να κλείσω αυτόν τον πρόλογο, με την ευχή η ανάγνωση του συγγράμματος να αφήνει ένα αίσθημα χαράς, αντίστοιχο της ικανοποίησης που έφερε σε μένα η συγγραφή του, σε συνδυασμό με τη εκτέλεση της υποκείμενης εργασίας γενικότερα.

Παλιακάσης Κωνσταντίνος

Σημείωση:

Η αναφορά -σε ένα κείμενο- μικρών αποστροφών (εως λίγες λέξεις) από κείμενα τρίτων, για λόγους σχολιασμού, διευκρινήσεων ή παραλληλισμών (ιδίως χωρίς κερδοσκοπικό πνεύμα), σύμφωνα με την παγκόσμια πρακτική, δεν αποτελεί ούτε διαφήμιση, ούτε κλοπή ή κατάχρηση της πνευματικής περιουσίας τους.

Εξάλλου, η έγκριση μιας διδακτορικής διατριβής από Πανεπιστημιακή Σχολή δεν υποδηλώνει και αποδοχή της γνώμης του συγγραφέως.

Περίληψη

Σε αντίθεση με την οργάνωση της φυσικής δομής των πρωτεϊνών, που η γνωστή ιεραρχία με τα τριτοταγή (δομικές ενότητες) και δευτεροταγή στοιχεία (α-έλικες και β-πτυχωτά φύλλα) την κάνει απλή και κατανοητή, η γνώση που αφορά -πιθανές- σχέσεις της με την αμινοξική αλληλουχία είναι ελλειπής. Τέτοια γνώση αποκτά μεγαλύτερη αξία καθώς, την ίδια στιγμή που οι τεχνικές σε επίπεδο DNA παράγουν μαζικά αλληλουχίες, ο πειραματικός προσδιορισμός της φυσικής δομής μιας πρωτεΐνης παραμένει χρονοβόρος, ενώ η πρόβλεψή της, με ενεργειακούς υπολογισμούς στο ατομικό επίπεδο, είναι εξαιρετικά δυσχερής αν όχι αδύνατη. Άλλωστε η ειδικότητα των αλληλεπιδράσεων στην τελική δομή δεν καθορίζεται στο επίπεδο αυτό. Όμως στο τριτοταγές επίπεδο οι πρωτεΐνες οργανώνονται με βάση (λίγα) δομικά πρότυπα, που δεν αναφέρονται στις λεπτομέρειες κάθε πρωτεΐνης, αλλά περιγράφουν τρόπους, με τους οποίους τα δευτεροταγή στοιχεία μπορούν να διευθετηθούν σε θερμοδυναμικά σταθερές και λειτουργικά χρήσιμες διατάξεις. Τοπολογικά διακριτές θέσεις κλειδιά, που μπορούν να οριστούν στις διατάξεις αυτές, έχουν συγκεκριμένους περιορισμούς -κυρίως λόγω αλληλεπιδράσεων με θέσεις γειτονικές στο χώρο- και αναμένεται να είναι συμβατές με λίγους μόνο από τους αμινοξικούς τύπους. Καθώς οι πρωτεϊνικές οικογένειες, που ακολουθούν κάθε πρότυπο, δεν έχουν άλλη ομοιότητα εκτός από τη δομική, ένα σύνολο (πιθανών) προτιμήσεων τέτοιου είδους εκφράζει τα χαρακτηριστικά που πρέπει να έχει μια αλληλουχία, ώστε να είναι συμβατή με το πρότυπο, και είναι ανεξάρτητο από την λειτουργία και εξελικτική προέλευση των οικογενειών.

Σε μια προσπάθεια να διερευνηθούν τα παραπάνω, καταστρώθηκε μια σχετική ανάλυση για ένα απλό δομικό πρότυπο, το δεμάτι που αποτελείται από τέσσερις α-ελικες πακεταρισμένες αντι-παράλληλα, με αριστερή συστροφή, που η απλότητά του το έχει κάνει συνήθη πλατφόρμα σε πειράματα ανάλυσης και σχεδιασμού. Οι α-έλικες που το αποτελούν είναι αμφιπαθικές και μπορούν να περιγραφούν σαν μια επανάληψη επτά θέσεων $(abcdefg)_n$, από τις οποίες οι a και d είναι υδρόφοβες, οι b, c και f εκτεθειμένες στο διαλύτη, και οι e και g στα όρια του υδρόφοβου πυρήνα με τον εξωτερικό χώρο. Αν και το δείγμα αποτελούταν από πρωτεϊνικές οικογένειες ποικίλης λειτουργίας και προέλευσης (εξελικτικά), τα αποτελέσματα ήσαν σαφή: δεν πρόκειται "απλά" για την αναμενόμενη κατανομή, με τα υδρόφοβα κατάλοιπα κρυμμένα στο εσωτερικό και τα υδρόφιλα έξω, αλλά για μια έντονη ανακατανομή στα πλαίσια αυτά. Μάλιστα, κάθε θέση καταλαμβάνεται σε ποσοστό μεγαλύτερο από 50% από όχι περισσότερους από πέντε αμινοξικούς τύπους, ενώ το 40% των θέσεων a και d από μόλις δύο: λευκίνη και αλανίνη. Ενώ όμως η κατανομή της αλανίνης είναι πιο ομοιόμορφη στις επτά θέσεις, εκείνη της λευκίνης -αντίθετα- είναι έντονα ανομοιόμορφη. Η αίσθηση της ανακατανομής εσωτερικά (στο πλαίσιο του δεματιού) ενισχύεται από το γεγονός, ότι η σύσταση του δεματιού σε αμινοξικούς τύπους δεν διαφέρει πολύ από εκείνη των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικες· και πάντως διαφέρει πολύ λιγότερο, απ'ότι διαφέρει -αντίστοιχα- της κάθε θέσης χωριστά.

Τα τμήματα, που συνδέουν τις α-έλικες μεταξύ τους, είναι συνήθως μικρά σε μήκος και άρα έχουν περιορισμένο ρεπερτόριο διαμορφώσεων: συνδετικά τμήματα που παρεμβάλλουν δύο ή τρία

κατάλοιπα μεταξύ των α-ελίκων, προτιμούν (ή ίσως απαιτούν) συγκεκριμένες διαμορφώσεις, ανεξάρτητες από το αν στρέφουν την κύρια αλυσίδα δεξιά ή αριστερά σε σχέση με τον υδρόφοβο πυρήνα (κάτι που -όμως- οδηγεί σε κάποια πρότυπα αμινοξικών τύπων κατά μήκος τους). Από την άλλη, οι α-έλικες του δεματιού δείχνουν μια σαφή προτίμηση να ξεκινούν και να τελειώνουν σε συγκεκριμένες θέσεις a-g. Συγκεκριμένα, οι τέσσερις θέσεις που ακολουθούν το N-άκρο προτιμούν να είναι θέσεις a-b-c-d ή b-c-d-e, ενώ οι τελευταίες πριν το C-άκρο (αντίστοιχα) g-a-b-c ή a-b-c-d. Αυτές οι αντιστοιχίσεις διευκολύνουν την -τυχόν- σύνδεση των ελίκων με μικρού μήκους συνδετικά τμήματα, όμως και άκρα που συνδέονται με μακρύτερα συνδετικά τμήματα έχουν τις ίδιες προτιμήσεις. Μέσα από ένα ιδιαίτερο ταίριασμα, οι ίδιοι αμινοξικοί τύποι που είναι συμβατοί με αυτές τις αντιστοιχίες, ικανοποιούν και τις προτιμήσεις που -όπως είναι γνωστό από τη βιβλιογραφία- έχουν οι ακραίες στροφές των ελίκων για συγκεκριμένους αμινοξικούς τύπους.

Το σύνολο των παραπάνω δεδομένων (κατανομές ακτινικά και κατά μήκος των ελίκων, συμπεριφορά των συνδετικών τμημάτων) αποτελεί μέρος εξειδικευμένης γνώσης για το συγκεκριμένο πρότυπο. Γνώση του είδους μπορεί να διευκολύνει όλο το φάσμα των διαδικασιών σχεδιασμού σε πρωτεΐνες, από την αποτίμηση αποτελεσμάτων που έχουν μικρές εντοπισμένες αλλαγές, μέχρι το σχεδιασμό ενός ολόκληρου πρωτεϊνικού μορίου από το μηδέν. Προκειμένου να δειχτεί αυτό, περιγράφεται σε λεπτομέρεια ένα παράδειγμα, όπου χρησιμοποιείται το σύνολο των πινάκων των προηγουμένων κεφαλαίων. Με βάση τέτοια πληροφορία, δεν χρειάζεται να αναλωθούν ώρες λεπτομερούς εργασίας μπροστά από την οθόνη των γραφικών μόνο και μόνο για να οργανωθεί το γενικό πλαίσιο της δομής του μορίου· το τελευταίο σχεδιάζεται ταχύτατα, αφήνοντας για διευκρίνιση, σε επίπεδο αλληλεπίδρασης με τον άνθρωπο ή για συστηματικές συνδυαστικές προσεγγίσεις, μόνο μικρές λεπτομέρειες που δεν εμπίπτουν στο είδος της πληροφορίας των πινάκων. Η μελέτη ολοκληρώνεται με μια εκτενή διερεύνηση, γύρω από τη στατιστική σταθερότητα των συμπερασμάτων, το είδος των προβλέψεων δομής από αλληλουχία στο οποίο συνεισφέρουν και το σχήμα και τις συνθήκες κάτω από τις οποίες μπορεί να επιτευχθεί. Η διερεύνηση αυτή αποκαλύπτει την πραγματική φύση των αποτελεσμάτων, των οποίων η ελαστικότητα βρίσκει μια ιδιαίτερη ερμηνεία και από μια μοριακή εξελικτική σκοπιά.

Abstract

Hierachic organisation of protein structure makes it simple to describe and follow. Yet, the knowledge regarding any -potential- detailed relationships between sequence and structure is far from complete. This sort of knowledge finds value in the fact that, while DNA-based techniques have permitted massive sequence discovery, structure determination by experimental means remains slow, and prediction in energetic (thermodynamic) terms is extremely difficult -if not useless; not to mention that the specificity of the interactions observed at the folded molecule is not encoded at this level. However, at the tertiary structural level, each protein follows one of -relatively- very few structural motifs; this might provide a key to the case. Tertiary structural motifs do not refer to any exact details of any particular protein family, but -instead- they describe how the secondary structural parts can be arranged into a thermodynamically stable and functional whole. Key locations, which may be defined as topologically distinct positions on each scaffold, obey limitations -due to specific interactions with neighbouring positions in 3D space- which should allow compatibility to only a few amino acid types. Since the protein families, which implement each motif, share no similarity other than a common scaffold, the set of these preferences, taken as a whole, should describe the characteristics a sequence must have in order to be compatible with the motif and is irrelevant to origin or function. Do such preferences exist?

Seeking an answer, a relevant analysis was performed on a simple recurrent tertiary motif, the four-α-helical bundle. The four helices comprising the motif are packed in an up-and-down or all-anti-parallel fashion; among the two possible twists, the left-handed was chosen. Stability and structural simplicity have proven this motif an attractive system for analysis and design procedures. The four helices are amphipathic and can be described as a regular repetition of seven positions $(abcdefg)_n$. Positions a and d are buried and hence hydrophobic, b, c and f are exposed to the solvent whereas e and g lie on the boundary between the interior and the exterior. Despite the functional and evolutionary variety among the families considered, the results are clear cut: it is not just the expected distribution with hydrophobic residues inside and hydrophilic ones outside; instead, within this framework, an extensive redistribution occurs, where each position is occupied in more than 50% of the instances by no more than five amino acid types, with a and d in about 40% by only two: Leu and Ala. However, while the distribution of alanine is rather smooth over the seven positions, leucine appears more determined. The feeling of this redistribution across (but within) the bundle is reinforced by the fact that the composition of the bundle does not differ markedly from that of all-α proteins; and -sure- it differs much less than it does, when each of the seven positions is considered separately.

The interconnecting segments are usually short and hence they have a limited repertoire of conformations: when they are only two or three residues long, prefer (or perhaps demand) particular conformations. These are independent from whether they turn the head of the main chain to the right or to the left relative to the hydrophobic core; however this detail settles for specific patterns of residue types along the connecting segment. Moreover, the helices comprising the bundle, prefer to begin and end at particular positions a-g. More specifically, the four positions following the N-cap prefer to be a-b-c-d or b-c-d-e, whereas the last four just before the C-cap prefer g-a-b-c or a-b-c-d. These preferences facilitate short connections, although ends connected via longer segments show similar preferences. Thanks to a multiple fit, the same amino acid types, compatible with this framework, satisfy the preferences of the end helical turns, for particular amino acid types (known from the literature).

These data, taken together (distribution both around and along the helices, behaviour of the interconnecting segments) constitute part of specialised knowledge regarding the particular structural motif. Knowledge of the kind can facilitate the complete broad spectrum of protein design procedures, ranging from evaluation of data resulting from local modifications up to the de novo design of a complete protein molecule. To demonstrate this, an example is described in detail, which utilises the complete set of matrices and observations made in previous chapters. This way, the general scaffold is rapidly specified, except for details not covered by the nature of these matrices, which are best left to human intervention and/or brute force, systematic, combinatorial approaches. This work is completed after an extensive investigation, concerning the statistics of the conclusions drawn, as well as the kind of sequence-based structure prediction they can facilitate, along with manners available and conditions necessary for it. This investigation reveals the real nature of the results, the tolerance of which finds a neat explanation on molecular evolutionary grounds as well.

Περιεχόμενα

Συμβουλετική Επιτροπή - Εξεταστική Επιτροπή:	i
Πρόλογος:	ii
Ευχαριστίες:	vi
Περίληψη (στην Ελληνική):	viii
Abstract:	x
Γενική Εισαγωγή:	Εισ.1-23
Μέρος Α / Κεφάλαιο Ι:	Α.Ι.1-18
Μέρος Α / Κεφάλαιο ΙΙ:	Α.ΙΙ.1-30
Μέρος Β / Κεφάλαιο Ι:	Β.Ι.1-22
Μέρος Β / Κεφάλαιο ΙΙ:	Β.ΙΙ.1-20
Βιβλιογραφία	Βιβ.1-6
Ανεξάρτητη Εργασία:	1-19

“Η ζωή είναι σαν ένα κουτί με διάφορα σοκολατάκια·
ποτέ δεν ξέρεις τι θα σου τύχει”

(Σε ελεύθερη απόδοση από την ταινία “Forrest Gump

Γενική Εισαγωγή

Σκοπός της παρούσας μελέτης

Η εργασία που ακολουθεί αφορά την σχέση ανάμεσα στη (γραμμική και επομένως μονοδιάστατατη) αμινοξική αλληλουχία και την (τρισδιάστατη) δομή των πρωτεϊνών. Πιο συγκεκριμένα, γίνεται μια προσπάθεια να βρεθεί αν υπάρχουν τρόποι, με τους οποίους η διευθέτηση των αμινοξικών καταλοίπων μιας πρωτεΐνης στο χώρο αντανακλάται στην (και ίσως μπορεί να προβλεφθεί από την) κατανομή των αμινοξικών καταλοίπων στην αλληλουχία της.

Στα πλαίσια της προσπάθειας αυτής, αναλύονται προτιμήσεις τοπολογικά διακριτών θέσεων ενός απλού δομικού προτύπου, για τους είκοσι αμινοξικούς τύπους που απαντούν στις πρωτεΐνες. Επίσης ελέγχεται, αν οι προτιμήσεις αυτές είναι αρκετά καθαρές, ώστε να μπορεί να λεχθεί με σιγουριά αν μια υποψήφια αλληλουχία είναι συμβατή με το πρότυπο. Τα συμπεράσματα που προκύπτουν επηρεάζουν μεθόδους προβλεψης δομής και πρωτεϊνικού σχεδιασμού, ενώ αγγίζουν και εξελικτικά θέματα.

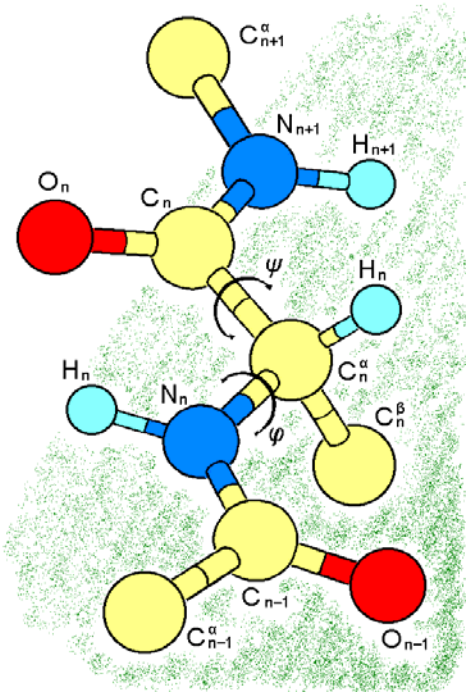
Στην "Εισαγωγή" που ακολουθεί, δίνονται κατά τρόπο όσο είναι δυνατό πλήρη αλλά και συμπαγή, οι ορισμοί των εννοιών που εμφανίζονται στην παρούσα μελέτη, καθώς και τα χαρακτηριστικά της δομής των πρωτεϊνών, που αφορούν στην εξήγηση των αποτελεσμάτων που ακολουθούν σε επόμενα κεφάλαια. Μια λεπτομερής και σε βασικό επίπεδο περιγραφή της δομής των πρωτεϊνών είναι αδύνατη στο παρόν πλαίσιο, και πιθανώς εκτός σκοπού. Για μια εξαιρετική περιγραφή στο επίπεδο αυτό, ο αναγνώστης μπορεί να αναφερθεί στο άρθρο ανασκόπησης των Richardson και Richardson [1989]. Επίσης, εξηγείται το πνεύμα της ανάλυσης, που ακολουθεί σε επόμενα κεφάλαια, το είδος της πληροφορίας που ζητείται, καθώς και η αναγκαιότητά της. Τέλος, παρουσιάζεται το πρότυπο που επελέγη για ανάλυση, το αντι-παράλληλο δεμάτι από τέσσερις α-έλικες με αριστερή συστροφή.

Βασικά χαρακτηριστικά της δομής των πρωτεϊνών-Ορισμοί

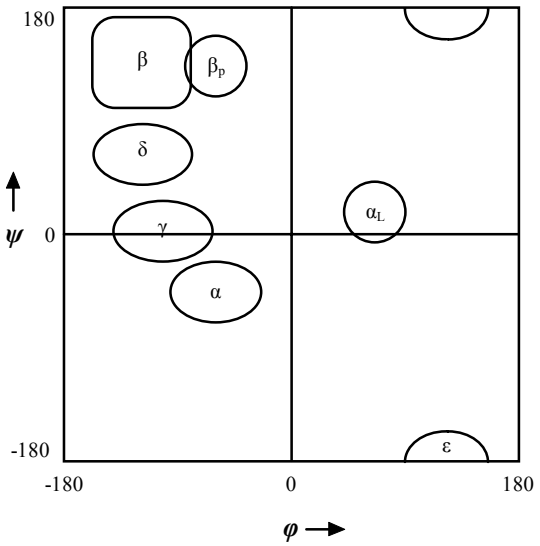
Οι πρωτεΐνες αποτελούν τον κύριο μηχανισμό υλοποίησης του γενετικού σχεδίου, που φέρουν τα μόρια των νουκλεϊκών οξέων. Αποτελούνται από (μία ή περισσότερες) πολυπεπτιδικές αλυσίδες, που αν και περιορίζονται σε ένα ρεπερτόριο είκοσι αμινοξικών τύπων καλύπτουν ένα μεγάλο αριθμό λειτουργιών που είναι αναγκαίες για τη ζωή.

Στη λειτουργική τους κατάσταση, παρατηρείται μια καλά καθορισμένη διευθέτηση των αμινοξικών καταλοίπων στο χώρο, που στο σύνολό της αποτελεί την *φυσική (native)* τρισδιάστατη δομή του μορίου. Σ'αυτή τη φυσική δομή, μέρος της αλληλουχίας (περίπου τα μισά αμινοξικά κατάλοιπα [Kabsch και Sander, 1983]) συμμετέχει στις γνωστές από την βασική βιοχημεία α-έλικες και β-πτυχωτές επιφάνειες (που συνήθως αναφέρονται και σαν *κανονικές* ή *περιοδικές δομές*), ενώ ένα μέρος συμμετέχει σε μη περιοδικές αλλά εξίσου σταθερές κατασκευές (*σύντομες αναστροφές (tight turns)* που περιλαμβάνουν *β-στροφές (β-turns)*, *σύντομες συνδέσεις (short straps)* και *συμπαγείς βρόγχους (compact loops)* [Chou και Fasman, 1977· Richardson και Richardson, 1989]). Τέλος, όχι σπάνια, ένα μέρος κατανέμεται σε τμήματα που δεν παρουσιάζουν συγκεκριμένη δομή, ούτε όταν η πρωτεΐνη κρυσταλλώσει.

Από πλευράς περιγραφής της φυσικής δομής, η τοπική διαμόρφωση κάθε αμινοξικού καταλοίπου εκφράζεται από τις γωνίες περιστροφής γύρω από δύο από τους απλούς δεσμούς που περιέχει (Εικόνα 1): την (δίεδρη) *γωνία φ*, που εκφράζει την περιστροφή γύρω από το δεσμό NH-C^{α}H, και τη (δίεδρη) *γωνία ψ*, που εκφράζει την περιστροφή γύρω από το δεσμό HC^{α}-CO. Οι γωνίες αυτές θεωρούνται 0^{O} όταν φέρνουν σε θέση *cis* το αμέσως προηγούμενο και το αμέσως επόμενο άτομο κατά μήκος της κύριας αλυσίδας, και 180^{O} όταν τα φέρνουν σε θέση *trans*. Αμινοξικά κατάλοιπα σε α-έλικες έχουν γωνίες $(\varphi,\psi) \approx (-60,-40)$, ενώ στα β-πτυχωτά φύλλα έχουν γωνίες $(\varphi,\psi) \approx (-120,+140)$. Η γωνία περιστροφής γύρω από τον πεπτιδικό δεσμό ονομάζεται *γωνία ω*, ενώ οι επί των πλευρικών αλυσίδων χαρακτηρίζονται σαν χ1, χ2 κοκ.



Εικ. 1. *Σχηματική αναπαράσταση ενός αμινοξικού καταλοίπου, μαζί με τμήματα των γειτονικών του κατά μήκος της αλληλουχίας. Επεξηγείται ο ορισμός των δίεδρων γωνιών φ και ψ ως περιστροφή γύρω από τους αντίστοιχους απλούς δεσμούς.*



Εικ. 2. *Διάγραμμα Ramachandran όπου δείχνονται οι στερεοχημικά επιτρεπτοί συνδυασμοί γωνιών (φ,ψ) για τα $_L$-αμινοξικά κατάλοιπα.*

Για εποπτικούς λόγους, τα ζεύγη των γωνιών αυτών απεικονίζονται σε ένα διδιάστατο διάγραμμα Ramachandran, (Εικόνα 2) με τις δίεδρες αυτές γωνίες σαν άξονες [Ramachandran et al, 1963]. Για λόγους στερεοχημικούς δεν είναι δυνατοί όλοι οι συνδυασμοί *(φ,ψ)*. Ορισμένες από τις περιοχές του διαγράμματος, που αντιστοιχούν σε επιτρεπτούς συνδυασμούς, συχνά αναφέρονται με ονόματα από τις αντίστοιχες β'-ταγείς δομές στις οποίες οδηγεί η επανάληψή τους (όλα τα θέματα ονοματολογίας ατόμων, δεσμών, γωνιών και δομών καλύπτονται από την IUPAC [1970]).

Η φυσική δομή των πρωτεϊνών είναι εξαιρετικά συμπαγής (με πυκνότητα και συμπιεστότητα συγκρίσιμη με εκείνη στερεών όπως ο πάγος [Dill, 1990 και εκεί αναφορές]), διαταράσσεται όμως εύκολα από αύξηση της θερμοκρασίας, μείωση ή αύξηση του pH, διαλύτες κ.α. Η πρωτεΐνη τότε, *χωρίς διάσπαση (ή δημιουργία) ομοιοπολικών δεσμών,* μεταπίπτει σε μια κατάσταση χωρίς καθορισμένη πλέον τρισδιάστατη δομή. Η διαδικασία της μετάπτωσης αυτής ονομάζεται *αποδιάταξη*. Η αντίστροφη διαδικασία, δηλαδή η απόκτηση της φυσικής δομής ξεκινώντας από μία τέτοια μη καθορισμένη τρισδιάστατη διαμόρφωση (μία από πολλές πιθανές) ονομάζεται *δίπλωμα (folding)*.

Σαν αποτέλεσμα της διαδικασίας του διπλώματος, κάποια τμήματα της αλληλουχίας έρχονται κοντά στο χώρο, δημιουργώντας πολλές επαφές μεταξύ τους, ενώ έχουν πολύ λιγότερες με τα υπόλοιπα τμήματα της αλληλουχίας. Αυτά συγκροτούν τις *δομικές ενότητες (structural domains)* [Janin και Chothia, 1985]. Βασικό χαρακτηριστικό των δομικών ενοτήτων των σφαιρικών υδατοδιαλυτών πρωτεϊνών είναι η απόκρυψη των υδρόφοβων πλευρικών αλυσίδων στο εσωτερικό, μακριά από το νερό, και η έκθεση των πολικών ή φορτισμένων στην επιφάνεια [δες για παράδειγμα Janin, 1979· Lesser et al, 1987].

Καθώς όσες πλευρικές αλυσίδες μπορούν να αλληλεπιδράσουν με το νερό βρίσκονται στην επιφάνεια της πρωτεΐνης, συχνά το εσωτερικό των πρωτεϊνών χαρακτηρίζεται σαν σταγόνα λαδιού, που περιβάλλεται από ένα υδρόφιλο κέλυφος που

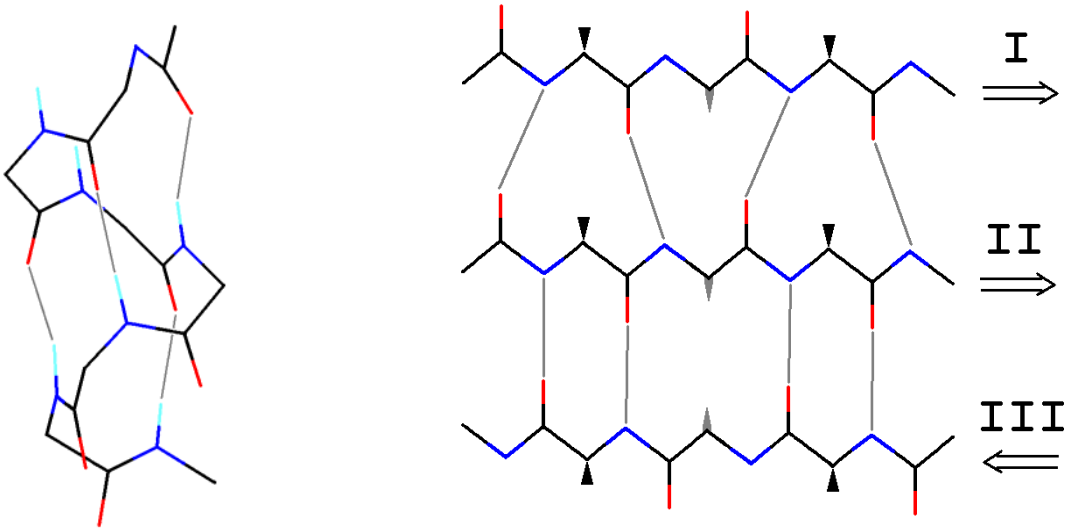
την κάνει διαλυτή (κάτι σαν την μεμβρανική διπλοστιβάδα, αλλά σε τρεις διαστάσεις). Αυτό δεν είναι ακριβές, γιατί οι πλευρικές αλυσίδες των αμινοξικών καταλοίπων του υδρόφοβου πυρήνα δεν έχουν ελευθερία κινήσεως όπως τα μόρια ενός υγρού ή στην επιφάνεια μιας μεμβράνης, αλλά απλά ταλαντώνονται γύρω από μια θέση ισορροπίας[2]. Από την άλλη πλευρά όμως, δεν πρόκειται για μια δομή στατική[3], "παγωμένη". Πέρα από τη συνεχή περιστροφή των -τυχόν- ευκίνητων πλευρικών αλυσίδων στην επιφάνεια του μορίου, τμήματα της κύριας αλυσίδας μπορούν να είναι ευκίνητα (ιδίως τα συνδετικά τμήματα μεταξύ των β'-ταγών στοιχείων, καθώς και τα αμινο- και καρβοξυτελικά άκρα), ενώ -όπως είναι γνωστό από τη βασική βιοχημεία- ολόκληρες δομικές ενότητες μπορούν να παραμορφώνονται ή/και να αλλάζουν σχετικό προσανατολισμό μεταξύ τους, ειδικά προκειμένου να εκτελέσει η πρωτεΐνη τη βιολογική λειτουργία της.

Το δεύτερο χαρακτηριστικό της οργάνωσης των πρωτεϊνών, που προκαλεί εντύπωση, είναι το πλήθος των ενδομοριακών δεσμών υδρογόνου [Baker και Hubbard, 1984]. Κάθε κατάλοιπο μπορεί να λειτουργήσει τόσο σαν δότης ενός τέτοιου δεσμού, δυνάμει του αζώτου του πεπτιδικού δεσμού, όσο και σαν δέκτης, μέσω του αντίστοιχου οξυγόνου του. Σε αυτούς πρέπει να προστεθούν εκείνοι τους οποίους μπορεί να σχηματίσει με πιθανές πολικές ομάδες της πλευρικής αλυσίδας του.

Ενεργειακοί υπολογισμοί στο *"διπεπτίδιο αλανίνης"* (δηλαδή ένα κατάλοιπο αλανίνης με μια πεπτιδική ομάδα από κάθε πλευρά: $CH_3C(O)NHC^{\alpha}H(CH_3)C(O)\ NHCH_3$ -Εικόνα 1), στους οποίους λαμβάνονται υπ'όψη μόνο δεδομένα στερεοχημικής παρεμπόδισης [Anderson και Hermans, 1988], δείχνουν ότι οι διαμορφώσεις εκείνες που είναι στερεοχημικά επιτρεπτές (στα μεμονωμένα κατάλοιπα), είναι εκείνες που, όταν επαναληφθούν σε διαδοχικά κατάλοιπα, οδηγούν σε β'-ταγείς δομές με ικανοποιημένους -ουσιαστικά- τους περισσότερους από τους πιθανούς δεσμούς υδρογόνου και με καλή γεωμετρία. Ομαλά, επαναλαμβανόμενα πρότυπα δεσμών υδρογόνου, μαζί με επαναλαμβανόμενες τοπικές διαμορφώσεις (όπως αυτές εκφράζονται από τις γωνίες περιστροφής *(φ,ψ)* γύρω από τους απλούς δεσμούς), σε διαδοχικά κατάλοιπα, ορίζουν αυτές που θεωρούμε "ομαλές" ("κανονικές", "περιοδικές") β'-ταγείς δομές [IUPAC, 1970] (Εικόνα 3).

[2] Μιλώντας αυστηρά, η "φυσική δομή" δεν είναι μια καθορισμένη μικροκατάσταση αλλά παριστά το σύνολο των (σχετικών) θέσεων ισορροπίας γύρω από τις οποίες γίνεται η ταλάντωση κάθε ατόμου· αν προστεθούν και -τυχόν- ελεύθερες περιστροφές πλευρικών αλυσίδων της επιφάνειας, τότε μιλάμε για μια συλλογή από τοπικά ελάχιστα, "γύρω" από τη "φυσική δομή" (στον πολυδιάστατο χώρο των βαθμών ελευθερίας του μορίου).

[3] Γενικά, μπορεί κανείς να διακρίνει τρεις χρονικές συνιστώσες στη δυναμική της δομής των πρωτεϊνών: (α) τη συνεχή ατομική κίνηση γύρω από καταστάσεις ισορροπίας, που μελετάται με προσομοιώσεις μοριακής δυναμικής (molecular dynamics) σε χρονικές κλίμακες από fs ως ns (β) το δίπλωμά τους, που τυπικά εκτυλίσσεται μέσα σε ms ως λίγα sec και μελετάται από την δομική βιολογία και την βιοχημεία και (γ) την εξέλιξή τους μέσα από τα δισεκατομμύρια έτη της ιστορίας τους, που μελετάται από την μοριακή και την εξελικτική βιολογία.



Εικ. 3. *Σχηματική αναπαράσταση των κυριότερων δομικών στοιχείων των πρωτεϊνών. Αριστερά: α-έλικα, Δεξιά: β-πτυχωτό φύλλο, όπου οι κλώνοι I και II είναι παράλληλοι ενώ οι II και III αντιπαράλληλοι (αν και συνήθως τα β-πτυχωτά φύλλα είναι αμιγή). Τα μαύρα τρίγωνα αντιστοιχούν σε πλευρικές αλυσίδες πάνω από το επίπεδο του χαρτιού, ενώ τα γκρίζα -αντίστοιχα- κάτω.*

Ένα ακόμη χαρακτηριστικό των πρωτεϊνών είναι μια ασυμμετρία σε όλα τα επίπεδα, που προκύπτει από το γεγονός ότι συμμετέχουν μόνο $_{L}$-αμινοξέα, και όχι τα $_{D}$-εναντιομερή τους[4]. Αυτή η ασυμμετρία αφορά τον α-άνθρακα και καθρεπτίζεται στην ασυμμετρία στο διάγραμμα Ramachandran σε επίπεδο καταλοίπου, αλλά οδηγεί στο να είναι δεξιόστροφες οι α-έλικες, ή οι διάφορες κάμψεις ενός β-φύλλου, και σε ασυμμετρίες στο πακετάρισμα και στη συνδεσμολογία σε υπερ-δευτεροταγείς δομές (παράδειγμα: Sternberg και Thornton, [1977]).

Οι β'-ταγείς δομές, μαζί με τις δομικές ενότητες, την συνολική δομή που συνήθως αποκαλούμε γ'-ταγή δομή και ενδεχόμενα ανώτερα υπερμοριακά συμπλέγματα, συγκροτούν μια *ιεραρχία* στη δομή των πρωτεϊνών, όπως τη γνωρίζουμε από τη βασική βιοχημεία. Αυτό είναι -ίσως- και το βασικό στοιχείο, που κάνει την "συνήθη" περιγραφή, όπως έγινε και στις μέχρι εδώ σελίδες, να "δείχνει" τόσο απλή και κατανοητή.

> Κατά πόσον η οργάνωση σε δομές κάτω από το επίπεδο των δομικών ενότητων έχει κάποια φυσική σημασία (πχ αποτελούν πυρήνες έναρξης του διπλώματος), ή απλώς προκύπτει από την ανθρώπινη παρατήρηση της τελικής δομής, δεν είναι γνωστό. Αν και σε διαδικασίες σχεδιασμού πεπτιδίων -μοντέλλων, για τη πειραματική μελέτη διαφόρων πλευρών της δομής των πρωτεϊνών, τα πεπτίδια σχεδιάζονται να έχουν άκρα που να

συμπίπτουν με άκρα β'-ταγών στοιχείων, αυτό μπορεί να μην έχει καμιά αντιστοιχία με ό,τι συμβαίνει π.χ. κατά το δίπλωμα μιας πρωτεΐνης.

[4] Δες Brady L. και Dodson G. (1994) "Reflections on a peptide" *Nature* **368**, 692-693 για πιθανές χρήσεις των $_{D}$-εναντιομερών σε τεχνητές πρωτεΐνες και πεπτίδια.

Προβλήματα στη πρόγνωση της πρωτεϊνικής δομής από απλές αρχές

Η πρόοδος στις τεχνικές του ανασυνδυασμένου DNA έκανε τον προσδιορισμό της αλληλουχίας των πρωτεϊνών κατά πολύ μαζικότερο από τον προσδιορισμό της τρισδιάστατης δομής, που συνήθως είναι κρυσταλλογραφικός. Έτσι, η απόσταση ανάμεσα στον αριθμό των νέων αλληλουχιών και τον αριθμό των νέων δομών διαρκώς αυξάνει, ανάγοντας σε βασικό στόχο της μελέτης της τρισδιάστατης δομής την ανάπτυξη μεθόδων πρόβλεψής της από την αμινοξική αλληλουχία· ενώ ο σχεδιασμός πρωτεϊνών με νέες ιδιότητες παραμένει προς το παρόν σε δευτερεύουσα θέση, παρά το γεγονός ότι οι τεχνικές υλοποίησης, οποιασδήποτε νέας πρωτεΐνης σχεδιαστεί, υπάρχουν και είναι οι ίδιες (δηλαδή σε επίπεδο DNA).

Η ιεραρχική οργάνωση της δομής των πρωτεϊνών, όπως παρατέθηκε στις προηγούμενες σελίδες, αποτελεί το κύριο μέρος οποιασδήποτε αντίστοιχης περιγραφής, σε ανάλογο πνεύμα. Αν και ο κύριος λόγος παράθεσής της, στα πλαίσια της παρούσας, ήταν η παρουσίαση βασικών εννοιών, που απαιτούνται για την ανάγνωση και κατανόησή της, φαίνεται παράλληλα, ότι ***η οργάνωση της δομής των πρωτεϊνών είναι απλή και καλά κατανοητή. Αντίθετα, η γνώση που αφορά την κωδικοποίηση της δομής στην αλληλουχία, ήδη για τα επίπεδα της δευτεροταγούς και τριτοταγούς δομής, είναι γενικής φύσεως και ελλειπής.*** Για παράδειγμα, είναι γνωστό ότι διάφοροι αμινοξικοί τύποι δείχνουν κάποια προτίμηση για μια από τις τρεις "κλασσικές καταστάσεις", π.χ. η βαλίνη προτιμά τους β-κλώνους, το γλουταμικό τις α-έλικες (κυρίως στην πρώτη στροφή), ενώ η γλυκίνη και τα μικρά υδρόφιλα είναι οι συχνότεροι θαμώνες στα μεταξύ τους συνδετικά τμήματα [Chou και Fasman, 1978]. Όπως επίσης ότι η κατανομή των υδρόφοβων καταλοίπων σε ένα κομμάτι αλληλουχίας είναι ενδεικτική του περιβάλλοντός του [Kyte και Doolitle, 1982]: ένα τμήμα 22 καταλοίπων με υψηλή υδροφοβικότητα "προκαταβάλλει" για μια διαμεμβρανική α-έλικα, ενώ μια εναλλαγή υδρόφοβων-υδρόφιλων με περίοδο 3,5 κατάλοιπα συνήθως αντιστοιχεί σε μια αμφιπαθική α-έλικα, πακεταρισμένη από τη μια της πλευρά πάνω σε μια υδρόφοβη επιφάνεια. Όμως, υπολογιστικές διαδικασίες, όπου επιχειρείται πρόβλεψη πρωτεϊνικής δομής από αλληλουχία, με βάση πληροφορία του είδους, έχουν -όπως εξηγείται και παρακάτω- περιορισμένη επιτυχία.

Θα μπορούσε να ισχυριστεί κανείς ότι γνώση σ'αυτό το "υψηλό επίπεδο" δεν χρειάζεται· ότι -αφού οι νόμοι που διέπουν τις αλληλεπιδράσεις μεταξύ των διαφόρων τύπων ατόμων, σαν συνάρτηση των αποστάσεών τους, είναι γνωστοί και απλοί- με την πρόοδο υπολογιστών και αλγορίθμων τελικά θα γίνει εφικτός ο υπολογισμός της δομής απευθείας, από απλές αρχές, από ένα "βασικό επίπεδο". Περιγράφονται στη συνέχεια εγγενή προβλήματα, που υπάρχουν στο βασικό αυτό επίπεδο, και που έχουν εμποδίσει σημαντικά τις μέχρι τώρα προσπάθειες για απευθείας πρόβλεψη της πρωτεϊνικής δομής από την αμινοξική αλληλουχία.

Αρχικά, καλό είναι να αναλύσει κανείς τη φυσική αιτία, που οδηγεί την πρωτεϊνική αλυσίδα στην ακολουθία των μεταβάσεων μεταξύ διαδοχικών πιθανών διαμορφώσεων στο χώρο, προκειμένου να μπορέσει να υλοποιήσει αλγορίθμους υπολογισμού του μεγέθους της. Από τη θερμοδυναμική

πλευρά, σε ένα σύστημα σταθερής σύνθεσης (δηλαδή χωρίς χημικές διεργασίες ή αλλαγή αριθμού σωματιδίων) και κάτω από σταθερή πίεση και θερμοκρασία (όπως προσεγγιστικά μπορεί να θεωρηθεί μια πρωτεΐνη που διπλώνει), η κατεύθυνση των διεργασιών είναι προς την πλευρά της *ελάττωσης της ελεύθερης ενέργειας Gibbs*[5] *(ΔG < 0).* Αν και εξαρτάται από τη θερμοκρασία[6] στην οποία μετριέται, τυπικές τιμές, για τη μετάπτωση από τη φυσική δομή στο συνοθύλευμα των αποδιαταγμένων διαμορφώσεων *(κατάσταση αποδιάταξης)*, είναι 5-15 Kcal/mole [Dill, 1990 και εκεί αναφορές]. Μάλιστα κινείται σε αυτά πλαίσια είτε πρόκειται για μια μικρή πεπτιδική ορμόνη είτε για μια μεγάλη δομική ενότητα αφού οι επιπλέον αλληλεπιδράσεις στις μεγάλες πρωτεΐνες αντισταθμίζονται από *αυξημένα εντροπικά φαινόμενα, που αποτελούν και την κύρια συνιστώσα που αντιτίθεται στο δίπλωμα*: στην κατάσταση αποδιάταξης, το κάθε αμινοξικό κατάλοιπο είναι ελεύθερο να περιστραφεί γύρω από τους απλούς δεσμούς που περιέχει, ενώ στη διπλωμένη δομή κάθε κατάλοιπο ακινητοποιείται σε μια από τις στερεοχημικά επιτρεπτές διαμορφώσεις, οδηγώντας σε εντροπικές απώλειες *τοπικά*, μέσα στα όρια του κάθε καταλοίπου[7]. Επίσης, στην κατάσταση αποδιάταξης μια πρωτεΐνη έχει στη διάθεσή της πολύ μεγαλύτερο χώρο, μέσα στον οποίο κινείται, και μέσα στον οποίο έχει πολύ περισσότερες δυνατές διαμορφώσεις, από ότι όταν διπλώσει και υποχρεωθεί σε μικρό όγκο, μέσα στον οποίο έχει πολύ λιγότερες δυνατότητες, λόγω του ότι δεν μπορεί να περάσει μέσα από τον εαυτό της, οδηγώντας πάλι σε απώλεια εντροπίας, αλλά *μη-τοπικά* αυτή τη φορά.

Θεωρητικά, λοιπόν, θα μπορούσε κανείς να εντοπίσει την φυσική δομή μιας πρωτεΐνης, αναζητώντας θερμοδυναμικά σταθερές δομές, στηριζόμενος σε βασικές αρχές της μηχανικής, αν και είναι ζωηρή ακόμη η συζήτηση, κατά πόσο η φυσική δομή αντιστοιχεί στο ολικό ελάχιστο της ελεύθερης ενέργειας, ή αποτελεί ένα τοπικό ελάχιστο προσπελάσιμο από ένα μεγάλο αριθμό αποδιαταγμένων διαμορφώσεων, σε χρόνο βιολογικά διαθέσιμο (δες Κεφ.Β.ΙΙ).

[5] Καθώς υπάρχει πληθώρα σχετικών συμβολισμών στη βιβλιογραφία, διευκρινίζεται ότι G=U+PV-TS, όπου P, V, T, S είναι οι γνωστές πίεση, όγκος, θερμοκρασία και εντροπία, ενώ U είναι η εσωτερική ενέργεια του συστήματος. Έτσι, αφού dU=TdS-PdV έπεται ότι dG=-SdT+VdP

[6] Φυσικά, για πρωτεΐνες που ακολουθούν ένα αμιγές μοντέλλο δύο καταστάσεων, δηλαδή φυσική-αποδιαταγμένη (N-U), στη θερμοκρασία αποδιάταξης T_m, είναι $\Delta G_{N-U}=0$: Αν P1 και P2 είναι η πιθανότητα με την οποία το σύστημα βρίσκεται στις καταστάσεις 1 και 2 που τις χωρίζει διαφορά ενέργειας ΔG, τότε P1/P2=exp(-βΔG) όπου β=α/T, T η απόλυτη θερμοκρασία και α εξαρτάται από τις μονάδες ενέργειας που χρησιμοποιούνται. Όταν P1=P2 έπεται ότι ΔG=0.

[7] Αξίζει να σημειωθεί εδώ ότι κατάλοιπα με μεγαλύτερη ευκολία περιστροφής γύρω από τις γωνίες *(φ,ψ)* όπως η γλυκίνη συνεισφέρουν περισσότερο στην κατάσταση αποδιάταξης, ενώ κατάλοιπα όπως η προλίνη, όταν αντικαθιστούν κατάλοιπα σε μια θέση όπου ταιριάζουν στερεοχημικά, προσφέρουν στην σταθερότητα και μόνο από το γεγονός ότι οδηγούν σε μικρότερη απώλεια εντροπίας.

> Πράγματι, για ένα μικρό μόριο, και παράγοντας όλες τις πιθανές διαμορφώσεις με τη βοήθεια ηλεκτρονικών υπολογιστών, μπορεί κανείς να υπολογίσει με ακρίβεια, όχι μόνο τη διαμόρφωση όπου η ελεύθερη ενέργεια φθάνει σε ολικό ελάχιστο, αλλά και την κινητική συμπεριφορά του μορίου, δηλαδή πως κατανέμει το χρόνο του στις εναλλακτικές διαμορφώσεις που αντιστοιχούν στα τοπικά ελάχιστα ενέργειας (όπως π.χ. Anderson και Hermans [1988]).

Όμως, για μια πρωτεΐνη, η συστηματική παραγωγή όλων των πιθανών διαμορφώσεων, με σκοπό τον εντοπισμό της φυσικής δομής με βάση ευνοϊκές αλληλεπιδράσεις, έχει αποκλεισθεί από νωρίς, για διάφορους λόγους. Κατ'αρχήν, η κύρια αλυσίδα μιας πρωτεΐνης 140 αμινοξικών καταλοίπων, μπορεί να περιστραφεί γύρω από περίπου 140 ζεύγη γωνιών *(φ,ψ)*. Έστω και μόνο με τις δύο επικρατέστερες τιμές για το κάθε ζεύγος (ελικοειδή και εκτεταμένη), έχουμε $2^{140} > 10^{42}$ πιθανές διαμορφώσεις, πριν καν ληφθούν υπ'όψη οι περιστροφές των πλευρικών αλυσίδων. Και φυσικά το παράδειγμα είναι ακραίο, αφού τα κατάλοιπα σε μια πρωτεΐνη, όχι μόνο καλύπτουν όλες τις περιοχές της Εικόνας 2, αλλά κάποια -λίγα- από αυτά βρίσκονται και εκτός, σε όχι ευνοούμενες διαμορφώσεις.

Ένα δεύτερο πρόβλημα, προκειμένου να εντοπίσει κανείς τη φυσική δομή με ενεργειακούς υπολογισμούς, ανακύπτει από την μικρή ενεργειακή διαφορά, που πειραματικά ευρίσκεται ότι χωρίζει την φυσική δομή από την κατάσταση αποδιάταξης, κάτι που άλλωστε εξηγεί την ευκολία με την οποία διαταράσσεται η φυσική δομή. Ακόμη και για μια μικρή δομική ενότητα, 5-15 Kcal/mole αναλογούν σε λιγότερο από 0,1 Kcal/mole για κάθε κατάλοιπο, όταν -συγκριτικά- $R*T$ =0,6 Kcal/mole είναι η ενέργεια που αντιστοιχεί στη θερμική κίνηση (όπου T=300^{O}K και R η σταθερά των αερίων 1,99 cal/(^{O}K*mole)). Πάντως, αυτό το φάσμα ενεργειών σταθεροποίησης επιλέχτηκε εξελικτικά, άρα μεγαλύτερες διαφορές ήσαν μάλλον βλαπτικές, (π.χ., ίσως δεν επέτρεπαν τις κινήσεις, που συνήθως απαιτούνται για την βιολογική λειτουργία μιας πρωτεΐνης). Λαμβάνοντας υπ'όψη τον διαλύτη, για τις περισσότερες αλληλεπιδράσεις, που παρατηρούνται στη φυσική δομή, υπάρχουν ισοδύναμες ή καλύτερες αλληλεπιδράσεις και όταν η πρωτεΐνη είναι σε κατάσταση αποδιάταξης: για παράδειγμα, οι ομοιοπολικοί δεσμοί είναι ίδιοι στις δύο καταστάσεις, ενώ και όλες οι πολικές ομάδες της πρωτεϊνικής αλυσίδας, που συμμετέχουν σε δεσμούς υδρογόνου στη φυσική δομή, προφανώς στην κατάσταση αποδιάταξης αλληλεπιδρούν π.χ. με τον διαλύτη. Κάθε αλληλεπίδραση που μεταβάλλεται, για να ευνοεί τη διπλωμένη μορφή, θα πρέπει να συνεισφέρει σε σταθερότητα αυξημένη έναντι της κατάστασης αποδιάταξης· έτσι, κάποιες μεταβολές την ευνοούν, και κάποιες όχι. Η (μικρή) ενέργεια σταθεροποίησης προκύπτει σαν το ισοζύγιο αυτών των μεταβολών, θετικών και αρνητικών. Έτσι, παρά τον μεγάλο αριθμό αλληλεπιδράσεων, που παρατηρείται στη διπλωμένη δομή, διαταραχή λίγων έστω από αυτές οδηγεί -συχνά- σε μια πρωτεΐνη που δεν διπλώνει.

Και φυσικά, τη φυσική δομή θα πρέπει κανείς να την ξεχωρίσει από έναν αριθμό διαμορφώσεων χαμηλής ενέργειας, από τις οποίες η διαφορά είναι ακόμη μικρότερη. Π.χ. Το N-ακετυλο-Ser-N'-μεθυλαμίδιο (ένα απλό κατάλοιπο σερίνης προστατευμένο από τις δύο πλευρές της κύριας αλυσίδας:

$CH_3C(O)NHC^{\alpha}H(OH)C(O)\ NHCH_3$) έχει 52 τοπικά ελάχιστα που απέχουν λιγότερο από 5 Kcal/mole από το ολικό [Purisima και Scheraga, 1986], ενώ ένα ολιγοπεπτίδιο μπορεί να έχει ήδη 100-1000 τοπικά ελάχιστα. Το δυσκολότερο: αυτές οι μικρές ενεργειακές διαφορές υπολογίζονται σαν άθροισμα ενός τεράστιου πλήθους μεταβολών αλληλεπιδράσεων. Ακόμη και οι ομοιοπολικοί δεσμοί, που είναι οι ίδιοι στις δύο καταστάσεις, αποτελούν όρο της ενέργειας αυτής, αφού μπορούν να επιμηκυνθούν ή να συμπιεστούν ελαφρά, προκειμένου να "ανακουφίσουν" άλλες χειρότερες τάσεις. Όμως, αν η επιθυμητή ακρίβεια, για ένα άθροισμα 10^6 όρων (παρομοίου μεγέθους), είναι της τάξης 1%, η ακρίβεια που απαιτείται για τους επιμέρους όρους είναι $\propto 1\%/\sqrt{10^6}=10^{-5}$, δηλαδή της τάξης των ppm.

Έτσι, ακόμη και αν η ταχύτητα των ηλεκτρονικών υπολογιστών αυξηθεί, τόσο ώστε να μπορούν να παράγουν ένα μεγάλο αριθμό από πιθανές διαμορφώσεις, οπότε -σε συνδυασμό με βελτίωση στους αλγορίθμους- να μπορεί να διερευνηθεί επαρκώς ο χώρος των διαμορφώσεων, οι προσεγγίσεις και οι παραδοχές, που -προς το παρόν- γίνονται στον υπολογισμό της συνεισφοράς των επιμέρους συνιστωσών αλληλεπιδράσεων, είναι αρκετά μεγάλες ώστε να κάνουν ανέφικτο τον θεωρητικό εντοπισμό της δομής ελάχιστης ενέργειας. *Ειδικά για τις σφαιρικές υδατοδιαλυτές πρωτεΐνες*, η μόνη δραστική διαφορά ανάμεσα στη διπλωμένη δομή και την κατάσταση αποδιάταξης, και που σημειώθηκε νωρίτερα σαν το βασικό τους χαρακτηριστικό, είναι η συγκέντρωση σε μεγάλο ποσοστό (>85% στην παρούσα ανάλυση) των υδρόφοβων καταλοίπων στο εσωτερικό, με έκθεση των υδρόφιλων στην επιφάνεια. Στην κατάσταση αποδιάταξης, τα μόρια του νερού κάνουν λιγότερους δεσμούς υδρογόνου όταν ακουμπούν σε υδρόφοβη επιφάνεια. Επιπλέον, η ακινητοποίηση σε σχέση με γειτονικά μόρια νερού, λόγω διάταξης γύρω από τις υδρόφοβες ομάδες -και άρα η απώλεια βαθμών ελευθερίας- οδηγεί σε εντροπικές απώλειες. Έτσι, με την απόκρυψη των υδρόφοβων, ώστε να είναι απροσπέλαστα από τον διαλύτη, κερδίζεται ενέργεια που έχει υπολογιστεί σε τιμές που ποικίλουν από 20 εως 80 cal/mole/$\square^2$ (υδρόφοβης επιφάνειας που απομακρύνεται από την επαφή με το νερό). Αυτή θεωρείται και ως η σημαντικότερη πηγή σταθερότητας των δομών των σφαιρικών υδατοδιαλυτών πρωτεϊνών [Dill, 1990]. Όμως, η ελλειπής κατανόηση της φύσης της υδροφοβικότητας και οι προσεγγίσεις που γίνονται για να εισαχθεί σαν όρος στη συνάρτηση της ελεύθερης ενέργειας, σε συνδυασμό με την ακρίβεια που απαιτείται, έχει δώσει -μέχρι τώρα τουλάχιστο- μάλλον πενιχρά αποτελέσματα, στην πρόβλεψη της φυσικής δομής από βασικές αρχές (δες Lins και Brasseur [1995] για μια προσπάθεια).

Ειδικότητα κατά το δίπλωμα

Ακόμη και αν γνωρίζαμε με μεγάλη ακρίβεια την συνάρτηση ελεύθερης ενέργειας και τις ιδιότητές της, γύρω από το ολικό και τα τοπικά ελάχιστα, θα μπορούσαμε (θεωρητικά) να εντοπίσουμε το ολικό ελάχιστο, αλλά το ενδιαφέρον θα περιορίζοταν σε ένα τεχνικό επίπεδο. Η φυσική δομή δεν είναι βέβαιο ότι αντιστοιχεί σε αυτό (περισσότερα στο Κεφ.Β.ΙΙ), ενώ είναι προφανές ότι οι πρωτεΐνες δεν δοκιμάζουν (συστηματικά ή μη) όλες τις πιθανές διαμορφώσεις, με θερμική περιστροφή γύρω από τους απλούς δεσμούς, μέχρι να βρουν τη σωστή. Αυτό έτσι κι'αλλιώς δεν μπορεί να γίνει λόγω του αστρονομικά μεγάλου αριθμού των πιθανών διαμορφώσεων: η πρωτεΐνη των 140 καταλοίπων, που αναφέρθηκε νωρίτερα σαν παράδειγμα, για να περάσει μία φορά από καθεμία από τις 10^{42} πιθανές διαμορφώσεις, και για 10^{-12} sec, θα ήθελε περισσότερο από 10^{30} sec δηλ. 10^{22} χρόνια. Ακόμη και δειγματοληψία του 10^{-10} των πιθανών διαμορφώσεων δίνει χρόνους μεγαλύτερους από την ηλικία του σύμπαντος.

Υπό αυτό το πρίσμα, πιο ρεαλιστική και περισσότερα υποσχόμενη είναι η προσπάθεια να μελετηθούν -υπολογιστικά πάντα- οι κινήσεις μιας πρωτεΐνης, είτε στη φυσική δομή, είτε καθώς διπλώνει *(προσομοίωση μοριακής δυναμικής - molecular dynamics (MD) simulation)*[8]. Η θέση του κάθε ατόμου, για κάθε "επόμενη" χρονική στιγμή, σε μικρά βήματα μέσα στο χρόνο (της τάξης του 1 fs = 10^{-15}sec), υπολογίζεται με βάση την ταχύτητα που έχει και τις δυνάμεις που του ασκούν τα υπόλοιπα άτομα. Καθώς όμως οι πιθανές διαμορφώσεις για τις πρωτεΐνες είναι αστρονομικά πολλές, η διαδικασία πρέπει να αφεθεί να λειτουργήσει για ένα μεγάλο αριθμό χρονικών βημάτων, προκειμένου να μπορέσουν να εξαχθούν συμπεράσματα, ενώ -παράλληλα- ο αριθμός των αλληλεπιδράσεων που πρέπει να υπολογιστούν για κάθε βήμα είναι μεγάλος. Έτσι, με τις καλύτερες υπολογιστικές τεχνικές, οι προσομοιώσεις δεν υπερβαί-νουν την κλίμακα των ns (τυπικά μερικές δεκάδες ps)[9], πολλές τάξεις μεγέθους κάτω από τα λίγα μs, όπου κινείται το -ταχύτερο- δίπλωμα. Γεγονότα, όπως η προσθήκη ενός καταλοίπου σε μια υπό επέκταση α-έλικα, που απαιτούν από 50ps-2ns, είναι -προς το παρόν- μόλις στο όριο τους [Thompson et. al, 2000], ενώ απρόσιτα -προς το παρόν- παραμένουν πιο αργά γεγονότα, όπως ο σχηματισμός ενός αρχικού "πυρήνα", από τον οποίο θα γίνει η επέκταση αυτή. Τεχνικά προβλήματα, όπως η σωστή χρήση των ηλεκτροστατικών δυνάμεων, υπάρχουν ακόμη και σήμερα [Sagui και Darden, 1999]· και σε κάθε περίπτωση, η πιθανότητα να συνεισφέρουν στην πρόγνωση της δομής, συνδέεται με την κατανόηση των ιδιοτήτων της κατάστασης αποδιάταξης, από την οποία θα πρέπει να ξεκινούν, που όμως ακόμα παραμένει ελλειπής.

> Όταν εξαντλούνται τα υπολογιστικά και αλγοριθμικά περιθώρια, στα πλαίσια των επιτρεπτών από τη Φυσική προσεγγίσεων (πχ χρήση πολλαπλών χρονικών βημάτων

[8] Με τεχνικές που έχουν εδώ τη βάση τους, μπορούν να υπολογιστούν και τα θερμοδυναμικά μεγέθη που σχετίζονται με τροποποιήσεις στο σύστημα (πχ μια μετάλλαξη). Δες Yun R.H. και Hermans J. (1991) *Protein Engineering* **4**, 761-766 για ένα παράδειγμα.

[9] Στα πλεονεκτήματα των προσομοιώσεων αυτών είναι ότι δίνουν μια λεπτομερέστατη εικόνα των γεγονότων σε ατομικό επίπεδο, ενώ σε περιπτώσεις σχεδιασμού πρωτεϊνών μπορούν να αποκαλύψουν πιθανά αδύνατα σημεία .

[Schlick et al, 1997]), ο πλέον κοινός διαθέσιμος τρόπος να επιταχύνει κανείς την προσομοίωση είναι να χρησιμοποιήσει υψηλή (προσομοιούμενη) θερμοκρασία, που όμως αντιστοιχεί σε συνθήκες αποδιάταξης[10]. Έτσι, συνήθως σε τέτοιες μελέτες, ξεκινά κανείς από μια φυσική δομή και προσομοιώνει τα πρώτα γεγονότα αποδιάταξης, καθώς η πρωτεΐνη θερμαίνεται, ελπίζοντας ότι αντιστοιχούν στα τελευταία γεγόνοτα του διπλώματος [Finkelstein, 1997]. Αυτή η προσέγγιση δεν πρέπει να συγχέεται με τη χρήση υψηλής (προσομοιούμενης) θερμοκρασίας για την υπέρβαση ενεργειακών εμποδίων σε εξειδικευμένες χρήσεις, όπως η βελτιστοποίηση της ερμηνείας των δεδομένων περίθλασης των κρυστάλλων από τους κρυσταλλογράφους.

Όμως, αξίζει να προσεχτεί, ότι *γενικά* αυτό το σκεπτικό δεν αναζητά το ολικό ενεργειακό ελάχιστο (γι'αυτό δεν αναφέρθηκε στην προηγούμενη συζήτηση, για ενεργειακούς υπολογισμούς), αλλά "αναμένει" ότι η φυσική δομή θα ανακύψει σαν αποτέλεσμα των κινήσεων της πρωτεΐνης, κάτι που ενδέχεται να πετύχουν αυτές οι διαδικασίες, όταν φτάσουν στην ωριμότητά τους· όμως, αυτό δεν αποτελεί -σε καμία περίπτωση- "απευθείας" υπολογισμό της φυσικής δομής από την αλληλουχία.

Καθώς η ελάττωση της ελεύθερης ενέργειας "σπρώχνει" την πρωτεΐνη -απλώς- μακριά από την κατάσταση αποδιάταξης, μια άλλη φυσική αιτία (για την οποία συζητώνται κάποιες υποθέσεις στο Κεφ. Β.ΙΙ) την "οδηγεί" -μέσα από τις διαδοχικές πιθανές διαμορφώσεις- όχι σε *οποιαδήποτε* συμπαγή δομή (οπότε θα επρόκειτο για τυχαία κατάρρευση), αλλά στη φυσική της δομή, στο χρόνο που είναι βιολογικά διαθέσιμος, προσδίδοντας έτσι μια *ειδικότητα στο δίπλωμα*. Μια φυσική αλληλουχία διπλώνει στην ίδια δομή, είτε συντίθεται στο ριβόσωμα, είτε με μεθόδους συνθετικής χημείας (για μικρές πρωτεΐνες φυσικά), όπου η σύνθεση προχωρά "ανάποδα" (από το καρβοξυτελικό άκρο προς το αμινοτελικό), είτε από συστήματα έκφρασης *in vitro* μετά από κλωνοποίηση, είτε ημισυνθετικά με συμπύκνωση.

Επιπλέον, στη φυσική δομή, δεν αλληλεπιδρά, λόγου χάρη, κάθε υδρόφοβη επιφάνεια *τυχαία* με κάποια άλλη, απλώς επίσης υδρόφοβη· αλλά κάθε αμινοξικό κατάλοιπο αλληλεπιδρά με τα κατάλοιπα εκείνα που "πρέπει" (Εικόνα 4). Εμφανίζεται δηλαδή και *ειδικότητα στις αλληλεπιδράσεις*. Αποτελεί δε *"ίδιον"* της κάθε (φυσικής) δομής σε τέτοιο βαθμό, ώστε -όπως είναι η κοινή εργαστηριακή εμπειρία- καθώς μια πρωτεϊνική αλυσίδα επιβαρύνεται με όλο και περισσότερες -τυχαίες[11]- αποσταθεροποιητικές μεταλλαγές, και υποθέτοντας ένα αμιγές μοντέλλο δύο καταστάσεων σε ισορροπία, ο κανόνας είναι ότι: ή διπλώνει στη φυσική δομή (με όλο και ελαττούμενη σταθερότητα) ή δεν διπλώνει καθόλου· δεν διπλώνει στη δομή μιας άλλης πρωτεΐνης. Μάλιστα αν υποτεθεί ότι μια πρωτεΐνη επιβαρύνεται με τόσες μεταλλαγές που το $\Delta G_{\sigma\tau\alpha\theta\epsilon\rho\acute{o}\tau\eta\tau\alpha\varsigma}$ να είναι θετικό (μετατοπίζοντας την ισορροπία προς την

[10] Υπάρχουν κι άλλοι "βίαοι" τρόποι επιτάχυνσης. Δες Williams M.A., Thornton J.M., και Goodfellow J.M. [1997] *Protein Engineering* **10**, 895-903 για χρήση "διευκολυνόμενης ενυδάτωσης του υδρόφοβου πυρήνα".

[11] Είναι θεωρητικά δυνατό να κατευθύνει κανείς τις μεταλλάξεις ώστε να οδηγήσει μια πρωτεΐνη να διπλώσει με τρόπο διαφορετικό από το φυσικό της. Δες Gross M και Plaxco K.W. [1997] *Nature* **388**, 419-420

κατάσταση αποδιάταξης), για παράδειγμα +5.5 Kcal/mole τότε, το 1 μόριο κάθε περίπου 10^4 που θα συνεχίσει να παρατηρείται διπλωμένο[12], θα διπλώσει με το φυσικό τρόπο.

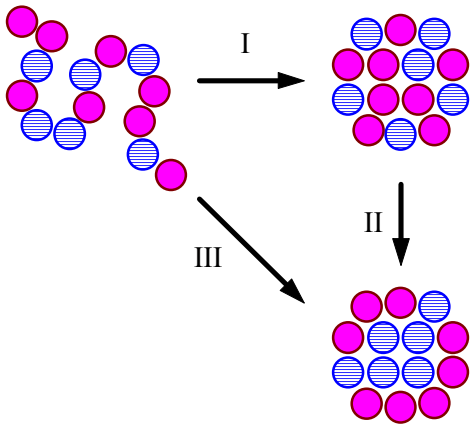
Όμως, παραμένει γεγονός, ότι οι ενδομοριακές αλληλεπιδράσεις, που παρατηρούνται στη φυσική δομή (δες και προσάρτημα στο τέλος του κεφαλαίου), δεν είναι ειδικές σε επίπεδο μεταξύ ατόμων ή αμινοξικών καταλοίπων. Οι δεσμοί υδρογόνου σε συνδυασμό με τη τοπική στερεοχημική παρεμπόδιση καθορίζουν επιτρεπτές τοπικές διαμορφώσεις, όχι όμως ποια από αυτές θα ακολουθήσει το κάθε κατάλοιπο. Επιπλέον, δεσμοί υδρογόνου μεταξύ α-ελίκων είναι σπάνιοι, ενώ στα β-πτυχωτά φύλλα, που φέρνουν κοντά τους επιμέρους β-κλώνους, μπορούν να γίνουν μεταξύ οποιωνδήποτε καταλοίπων (και όχι όπως π.χ. οι δισουλφιδικοί δεσμοί, που γίνονται μόνο μεταξύ κυστεϊνών). Έτσι, δεν μπορεί να συνάγει κανείς, αν θα φέρουν κοντά στο χώρο κάποια συγκεκριμένα κατάλοιπα, μακρινά στην αλληλουχία. Εξαίρεση αποτελούν ζεύγη φορτίων θαμμένα μέσα σε υδρόφοβα περιβάλλοντα, που συνήθως βοηθούν στην ειδικότητα -στο "σωστό ταίριασμα" ανάμεσα σε δύο τμήματα [DeGrado et al, 1999]- και αναφέρεται μια τέτοια περίπτωση στο Κεφ.Β.ΙΙ. Το πακετάρισμα είναι τόσο καλό, που το εσωτερικό των πρωτεϊνικών μορίων να έχει ιδιότητες στερεών, όμως έχει αποδειχτεί ότι μπορεί να επιτευχθεί γενικά εύκολα μεταξύ β'-ταγών στοιχείων [Behe et al, 1991], και επομένως δεν προσφέρει ειδικότητα. Η διαδικασία του διπλώματος οδηγεί σε μια κατανομή των πλευρικών αλυσίδων στο σωστό για την καθεμία περιβάλλον (και μπορεί να θεωρηθεί σε ένα βαθμό σαν μια τέτοια διαδικασία)· αυτή η κατανομή είναι επίσης που διαφοροποιεί το δίπλωμα των πρωτεϊνών από τις διαδικασίες ακανόνιστης κατάρρευσης ομοπολυμερών σε συμπαγείς δομές (Εικόνα 4). Όμως, αν και είναι η κύρια πηγή σταθερότητας της διπλωμένης δομής, σαν κατανομή δεν είναι απόλυτη, αφού πχ. πολλές από τις πλευρικές αλυσίδες στην επιφάνεια είναι υδρόφοβες (δες για παράδειγμα Κεφ.Α.Ι). Επομένως, ***αν υπάρχει ειδικότητα στην αλληλεπίδραση, είναι μεταξύ τμημάτων της πρωτεΐνης*** (όπως διδάσκεται στη βασική βιοχημεία, δηλαδή σαν ταίριασμα επιφανειών π.χ. β'-ταγών στοιχείων, με συμπληρωματικές ιδιότητες). ***Αποτελεί, δε, κύριο άξονα της παρούσας εργασίας, το ερώτημα αν η ειδικότητα -σε αυτό το επίπεδο- προδια-γράφεται στην αλληλουχία τέτοιων τμημάτων***, ενώ ο βαθμός στον οποίο περιλαμβάνει και κινητικά αποτελέσματα (δηλαδή κινήσεων της πρωτεΐνης, όπως περιγράφονται σε όρους μοριακής δυναμικής) μένει να δειχτεί.

[12] Εννοείται στατιστικά, σε μια οποιαδήποτε δεδομένη στιγμή, καθώς η ισορροπία είναι δυναμική και όχι στατική. Το ίδιο μόριο που τη μια στιγμή είναι διπλωμένο λίγο αργότερα μπορεί να έχει μεταβεί σε άλλη κατάσταση, ενώ κάποιο άλλο στο μεταξύ να έχει διπλώσει. Ο χρόνος που συμβαίνουν αυτά, εξαρτάται από τις συνθήκες κάτω από τις οποίες γίνεται η παρατήρηση και ασφαλώς από τη θερμοκρασία.

Ποια μορφή μπορεί να πάρει η αντανάκλαση στην αλληλουχία, της ειδικότητας της δομής και των αλληλεπιδράσεων σε επίπεδο μεγαλυτέρων τμημάτων; Πριν από την παρουσίαση της προσέγγισης που επιχειρείται στην παρούσα εργασία, περιγράφονται δύο ακραία πρότυπα κατανομής της πληροφορίας σε ένα μέσο, που βοηθούν στην κατανόηση του πως μπορεί να ταξινομηθεί τέτοια πληροφορία κατά μήκος μιας αλληλουχίας[13]:

(α) *Κεντρικά:* σε αυτή την περίπτωση η πληροφορία βρίσκεται σε ένα ή λίγα τμήματα της αλυσίδας. Στην πιο ακραία μορφή μάλιστα, κάθε τμήμα έχει ολόκληρο ένα κομμάτι της πληροφορίας (για το πως θα διπλώσει το ίδιο και πως θα αλληλεπιδρά με τα υπόλοιπα στη διπλωμένη μορφή), που δεν υπάρχει πουθενά αλλού στην αλληλουχία. Αυτό επιτρέπει να γίνεται το δίπλωμα ανεξάρτητα κατά ενότητες, όμως συνεπάγεται ότι κατάλοιπα έξω από τα τμήματα που κρατούν την πληροφορία μπορούν να παθαίνουν σοβαρές αλλαγές χωρίς το μόριο συνολικά να καταρρέει, ενώ ακόμη και μια μικρή αλλαγή μέσα στα τμήματα αυτά μπορεί να αποδειχθεί καταστροφική.

(β) *Κατανεμημένη:* Κάθε κατάλοιπο περιέχει (ανεξάρτητα) μέρος μιας συνολικής πληροφορίας· η πληροφορία που περιέχει, όμως καλύπτεται σε κάποιο βαθμό και από άλλα κατάλοιπα. Αν συμβεί μια αλλαγή, η πληροφορία που έχει το κατάλοιπο χάνεται εντελώς μόνο αν χαθούν και τα υπόλοιπα. Στην πιο ακραία μορφή κάθε κατάλοιπο έχει μέρος της πληροφορίας για όλη τη δομή, και έτσι καθώς, αλλάζουν κάποια κατάλοιπα, η σταθερότητα πέφτει σταδιακά και ανεξάρτητα από το ποια κατάλοιπα αλλάζουν. Εδώ απαιτείται μια απόλυτη συνεργατικότητα. Όλα τα κατάλοιπα της αλληλουχίας συμμετέχουν ανεξάρτητα στην ολική σταθερότητα. Αποσταθεροποιητικές αλλαγές μπορούν να συμβούν σε οποιοδήποτε κατάλοιπο, και λειτουργούν αθροιστικά.



Εικ. 4. *Το δίπλωμα των πρωτεϊνών (III) είναι μια διαδικασία διαφορετική από την τυχαία κατάρρευση των πολυμερών σε συμπαγείς δομές (I) αφού περιλαμβάνει και έντονα φαινόμενα ανα-κατανομών (II).*

[13] Για περισσότερες λεπτομέρειες ο αναγνώστης μπορεί να ανατρέξει σε οποιοδήποτε εισαγωγικό εγχειρίδιο θεωρίας της πληροφορίας, ενώ το θέμα καλύπτουν και βιβλία που σχετίζονται με τεχνητή ευφυΐα, νευρωνικά δίκτυα, οργάνωση τραπεζών πληροφοριών, τηλεπικοινωνιακών συστημάτων και άλλων σχετικών εφαρμογών.

Το κεντρικό πρότυπο "φαίνεται" -διαισθητικά- πιο κοντά στην γενικότερη ιεραρχική οργάνωση της δομής των πρωτεϊνών, ενώ το κατανεμημένο -αντίστοιχα- πιο κοντά σε -άπειρα- πειραματικά δεδομένα μεταλλαξογένεσης και στη συνεργατικότητα των κινήσεων της -διπλωμένης- πρωτεΐνης.

Πιθανώς η πραγματικότητα είναι ένας συγκερασμός των δύο προτύπων υπό κάποιες αναλογίες. Ο βαθμός της επιτυχίας των "κλασσικών" (ή στατιστικών) μεθόδων πρόβλεψης β'-ταγούς δομής είναι ίσως ενδεικτικός [Chou και Fasman, 1977-1978· Garnier και Robson, 1989]. Στις μεθόδους αυτές συχνά υπολογίζονται οι προτιμήσεις των 20 αμινοξικών τύπων (ή συνδυασμών) για συμμετοχή σε διάφορες δομές και διαμορφώσεις. Οι προτιμήσεις αυτές είναι το αποτελέσμα της συμβατότητας των φυσικοχημικών ιδιοτήτων της(ων) πλευρικής(ων) αλυσίδας(ων) με τις απαιτήσεις των κατηγοριών διαμορφώσεων για τις οποίες γίνεται πρόβλεψη. Σε άλλες πάλι επιχειρείται η εύρεση τοπικών χαρακτηριστικών, όπως μια κανονικότητα στην υδροφοβικότητα· εντοπίζονται έτσι πρώτα οι πιθανές α-έλικες και οι β-κλώνοι αναζητούνται στην υπόλοιπη αλληλουχία βάσει κανόνων [Lim, 1974α-β]. Το ποσοστό επιτυχίας των μεθόδων αυτών συνδυασμένων είναι περίπου 65%. Όπου ορίζονται κατηγορίες, είναι καλά διακριτές (δηλαδή με διαφορετική στατιστική συμπεριφορά η καθεμία) και αρκετά γενικές (δηλαδή με μεγάλη εκπροσώπηση ακόμα και σε σχετικά μικρά δείγματα). Έτσι, αύξηση του δείγματος δεν αύξησε την αποτελεσματικότητα των κλασικών μεθόδων πρόβλεψης[14]. Αυτό σημαίνει ότι το πρόβλημα δεν ήταν τα λιγοστά δεδομένα, αλλά το γεγονός ότι η συνολική δομή στην οποία συμμετέχει ένα κατάλοιπο ή ένα ολιγοπεπτίδιο επιδρά έντονα στην τοπική του διαμόρφωση. Έτσι, όπως είναι από παλιά γνωστό [Argos, 1987], μικρά ολιγοπεπτίδια (μήκους μέχρι 4-5 κατάλοιπα) με την ίδια αλληλουχία μπορούν να έχουν διαφορετική διαμόρφωση σε διαφορετικά δομικά περιβάλλοντα, αν και 1/5 από αυτά δείχνει *προτίμηση* για μια μόνο διαμόρφωση. Αυτό το μέγεθος επανεξετάζεται κατά καιρούς από άλλους ερευνητές, καθώς οι τράπεζες δεδομένων μεγαλώνουν. Προς το παρόν, ολιγοπεπτίδια μήκους 8-9 καταλοίπων, με ταυτόσημη αλληλουχία, *έχουν παρατηρηθεί* σε διαφορετική διαμόρφωση, χωρίς όμως να αποτελούν τον κανόνα. Η σταδιακή -δε- πτώση της συχνότητας καθώς αυξάνει το μήκος, είναι ενδεικτική συγκερασμού: αν ίσχυε ένα ακραίο συνεργατικό πρότυπο, οι στατιστικές μέθοδοι δεν θα είχαν καμιά επιτυχία, ενώ αντίστροφα, αν ίσχυε ένα ακραία κεντρικό πρότυπο, θα αρκούσαν οι προβλέψεις για κάθε κομμάτι της αλληλουχίας χωριστά· άλλωστε, είναι η συνεργατικότητα στη δόμηση που (για παράδειγμα) επιτρέπει -στη

[14]Μπορεί όμως να αυξήσει των αριθμό των κατηγοριών για τις οποίες γίνεται πρόβλεψη.

διπλωμένη πλέον πρωτεΐνη- τη διάδοση κινήσεων μέσα από τη δομή (πχ. αλλοστερικά φαινόμενα). Όμως απέχει από μια ακραία συνεργατικότητα, όπου όλα τα κατάλοιπα θα συμμετείχαν εξίσου, αφού ένα κατάλοιπο μπορεί να εκπληρώνει πολλαπλές υποχρεώσεις από την ίδια θέση, ή διαφορετικούς ρόλους σε διαφορετικά περιβάλλοντα, ενώ και η πειραματική πράξη (μέσα στην οποία και τα αποτελέσματα της παρούσης) δείχνει ότι ορισμένα κατάλοιπα είναι πιο σημαντικά για την ακεραιότητα της δομής από άλλα.

Τριτοταγή Δομικά Πρότυπα

Ένα τελευταίο σημείο, που συνήθως ολοκληρώνει γενικές περιγραφές της ιεραρχικά οργανωμένης δομής των πρωτεϊνών (σε βασικό επίπεδο), σαν εκείνη των πρώτων σελίδων του παρόντος κεφαλαίου, είναι η παλιά παρατήρηση ότι οι πρωτεΐνες διπλώνουν με βάση λίγα *τριτοταγή δομικά πρότυπα*.

Πρώτοι οι Levitt και Chothia [1976] έκαναν μια πρώτη ταξινόμηση 31 γνωστών δομών σε τέσσερις κατηγορίες: (α) αποτελούμενες αποκλειστικά από α-έλικες, (β) αποτελούμενες αποκλειστικά από β-πτυχωτά φύλλα, (γ) αποτελούμενες από χωριστές περιοχές πλούσιες σε α-έλικες ή β-πτυχωτά φύλλα και (δ) αποτελούμενες από εναλλαγές β'-ταγών στοιχείων, δηλαδή α-έλικες που εναλλάσσονται με β-κλώνους. Επίσης εισήγαγαν την έννοια της *μονάδας διπλώματος (folding unit)*, χωριστά από την έννοια της δομικής ενότητας. Πρόκειται για αθροίσματα από 2 ή 3 β'-ταγή στοιχεία που βρίσκονται κοντά στην αλληλουχία και πακετάρονται μαζί πχ η α-φουρκέττα, δηλαδή δύο α-έλικες, που πακετάρονται αντιπαράλληλα, ενώ στην αλληλουχία τις χωρίζουν λίγα αμινοξικά κατάλοιπα.

Καθώς όλο και περισσότερες δομές προσδιορίζονταν πειραματικά, γινόταν όλο και πιο φανερό ότι οι πρωτεϊνικές δομές διπλώνουν με βάση λίγα *(τριτοταγή) δομικά πρότυπα (tertiary structural motifs)*. Οι Richardson και Richardson [1989] αναφέρουν ήδη ότι, μεταξύ των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικες, πιο κοινές είναι αυτές που αποτελούνται από τέσσερις α-έλικες πακεταρισμένες αντι-παράλληλα σαν ένα *δεμάτι (bundle)*· ότι άλλα δεμάτια από περισσότερες α-έλικες υπάρχουν, αλλά είναι λιγότερο συχνά· και ότι η μόνη άλλη διάταξη που απαντά συχνά, είναι αυτή που αποκαλείται Greek key, με πολλές από τις έλικες πακεταρισμένες κάθετα μεταξύ τους. Όπως επίσης αναφέρουν σαν πιο κοινές περιπτώσεις, για τις άλλες κατηγορίες, τα *β-βαρέλια τύπου ανοσοσφαιρινών*, και το *$(βα)_8$ βαρέλι*, που συνήθως αποκαλείται *TIM-barrel*, από τον κύριο αντιπρόσωπο, την ισομεράση της φωσφορικής τριόζης (triose-phosphate isomerase, TIM).

Είναι φανερό ότι αυτά τα δομικά πρότυπα δεν αναφέρονται στις λεπτομέρειες της κάθε συγκεκριμένης δομής, αλλά περιγράφουν τρόπους, με τους οποίους, β'-ταγή στοιχεία μπορούν να πακεταριστούν σε θερμοδυναμικά σταθερές και λειτουργικά χρήσιμες διατάξεις. Αποτελούν δε μια πιο γενική κατάταξη από τις οικογένειες πρωτεϊνών, αφού οικογένειες διαφορετικές από πλευράς εξελικτικής προέλευσης και λειτουργίας μπορούν να διπλώνουν με τον ίδιο τρόπο. Έχει προταθεί [Chothia, 1992], ότι ο αριθμός των πρωτεϊνικών οικογενειών δεν υπερβαίνει τις 1000 (*υποθέτοντας όμως, ότι κοινή τριτοταγής δομή τις περισσότερες φορές σημαίνει κοινή προέλευση*), όμως ο αριθμός των δομικών προτύπων αναμένεται μικρότερος.

Οι Orengo et al [1993] σε μια συστηματική ανάλυση των πρωτεϊνικών μορίων με πειραματικά προσδιορισμένη μορφή, μέτρησαν 112 μη-ανάλογες *δομικές οικογένειες*. Όμως, τα δομικά πρότυπα που περιελάμβαναν ήσαν ακόμη λιγότερα, αφού δομικές οικογένειες με μεγαλύτερο μέγεθος, συχνά αποτελούνταν από ένα πρότυπο που είχε ήδη βρεθεί για κάποια μικρότερη δομική οικογένεια συν

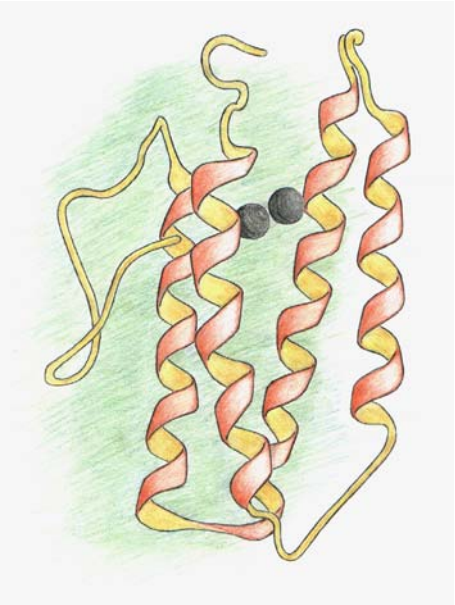
κάποια επιπλέον β'-ταγή στοιχεία. Οι ακριβείς αριθμοί υπολογίζονται ξανά κάθε τόσο, από διάφορους ερευνητές, και φυσικά εξαρτώνται από τα κριτήρια που θέτει ο καθένας κάθε φορά.

> Για παράδειγμα, οι Orengo et al [1993], αφού ελάττωσαν τις 1800 πρωτεϊνικές δομές, που είχαν κατατεθεί στην τρέχουσα έκδοση της τράπεζας δομικών πρωτεϊνικών δεδομένων Protein Data Bank [PDB· Bernstein et al, 1977], σε ένα αντιπροσωπευτικό σύνολο <300, απομακρύνοντας εκείνες για τις οποίες υπήρχαν ομόλογες (>35% ταυτότητα σε επίπεδο αλληλουχίας), αντιστοίχισαν στο κάθε κατάλοιπο κάθε πρωτεϊνικής δομής το σύνολο των διανυσμάτων προς τα β-άτομα άνθρακα όλων των υπολοίπων καταλοίπων στην ίδια δομή. Έπειτα, σε κάθε ζεύγος δομών υπό σύγκριση, αναζήτησαν τμήματα αλληλουχίας των οποίων τα κατάλοιπα είχαν κοινά περιβάλλοντα στις δύο πρωτεΐνες και ανέφεραν σαν μη-ανάλογα σε όσα η αντιστοίχιση ήταν κάτω από 70%. Για 80% το αποτέλεσμα αλλάζει σε 150.

Τα δομικά πρότυπα μπορούν να περιγραφούν σε όρους τοπολογικά διακριτών θέσεων που τα αποτελούν: κάθε τέτοια θέση ορίζεται έμμεσα, κυρίως από το τοπικό περιβάλλον που δημιουργεί η διευθέτηση των γειτονικών της θέσεων στον γύρο χώρο. Πχ α-έλικες, που πακετάρονται με τη μια πλευρά απέναντι από μια υδρόφοβη επιφάνεια, περιγράφονται σαν μια επανάληψη 7 θέσεων [Crick, 1953], αφού χρειάζονται περίπου 7 κατάλοιπα για να συμπληρω-θούν δύο στροφές α-έλικας. Αργότερα στη βιβλιογραφία, η επανάληψη αυτή συμβολίζεται σαν $(abcdefg)_n$, όπου οι θέσεις a και d αντικρύζουν τον υδρόφοβο πυρήνα, με τις b, c και f στην αντίθετη κατεύθυνση, ανεξάρτητα από το γενικότερο δομικό πλαίσιο στο οποίο παρατηρείται.

Το πακετάρισμα στο εκάστοτε περιβάλλον, και η ανάγκη για ευνοϊκή αλληλεπίδραση μαζί του, αναμένεται να επιτρέπει να είναι συμβατά με τη κάθε θέση λίγα μόνο από τα είκοσι αμινοξικά κατάλοιπα. ***Εαν όμως πρωτεΐνες που υιοθετούν ένα δομικό πρότυπο δεν παρουσιάζουν άλλη ομοιότητα (εκτός από τη δομική), τότε το σύνολο των προτιμήσεων αυτών θα πρέπει να αντιπροσωπεύει τα γενικά χαρακτηριστικά που πρέπει να ικανοποιεί μια πρωτεϊνική αλληλουχία για να είναι συμβατή με το πρότυπο αυτό, και είναι ανεξάρτητο από προέλευση και λειτουργία. Και φυσικά έχει το δυναμικό να οδηγήσει σε ειδικότητα στην αλληλεπίδραση,*** αφού θέσεις συνεχόμενες (ή με συγκεκριμένες αποστάσεις μεταξύ τους) σε ένα τμήμα αλληλουχίας, μπορούν -όταν διπλώσει το τμήμα αυτό- να δημιουργούν επιφάνειες, που μαζί με τις ιδιότητές τους προσφέρουν τη βάση για ειδική αναγνώριση στο πακετάρισμα.

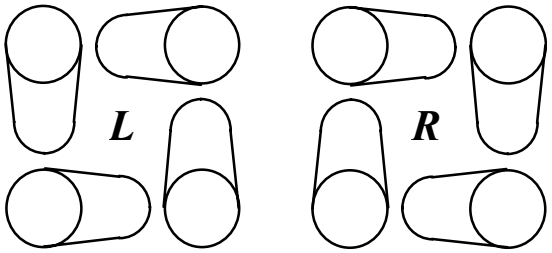
Υπάρχουν τέτοιες προτιμήσεις; Και αν ναι, μπορούν να κατατάξουν με ασφάλεια μια αλληλουχία στο δομικό πρότυπο με βάση το οποίο διπλώνει; Είναι ειδικές για κάθε δομικό πρότυπο, ή συναφή αλλά διαφορετικά δομικά πρότυπα μπορούν να έχουν παρόμοιες προτιμήσεις; Επιχειρείται στην παρούσα εργασία μια διερεύνηση των ερωτημάτων αυτών, με τη μορφή της ανάλυσης ενός δομικού προτύπου. Από διαφόρα πιθανά δεμάτια α-ελίκων, επιλέχθηκε εκείνο που αποτελείται από τέσσερις α-έλικες, με τις γειτονικές ανά δύο αντιπαράλληλες (Εικόνα 5). Αυτό το δομικό πρότυπο (στο οποίο θα αναφερόμαστε στο εξής απλά ως "δεμάτι"), αποτελεί τη βάση της οργάνωσης ενός ευρέος φάσματος πρωτεϊνικών οικογενειών, ποικίλης προέλευσης (εξελικτικά) και λειτουργίας.



Εικ. 5*. Σχηματική αναπαράσταση του μορίου μυοαιμερυθρίνη, ενός τυπικού αντιπροσώπου του δομικού προτύπου "δεμάτι τεσσάρων α-ελίκων" που χρησιμοποιήθηκε στην ανάλυση.*
(Οι μαύρες σφαίρες αντιστοιχούν σε άτομα σιδήρου.)

Οι βασικές τοπολογικές ιδιότητές του είχαν ήδη μελετηθεί σε έκταση όταν ξεκίνησε η παρούσα εργασία (το 1989) [Weber και Salemme, 1980· Banner *et al*, 1987· DeGrado *et al*, 1989· Presnel και Cohen 1989], όπως επίσης η θερμοδυναμική του σταθερότητα [Sheridan *et al*, 1982· Chou *et al*, 1988· Gilson και Honig, 1989], και εξακολουθούσαν να μελετώνται [Cohen και Parry, 1990· Carlacci και Chou, 1990α,β· Paliakasis και Kokkinidis, 1991, 1992· Chou *et al*, 1992· Chou και Zheng, 1992], ενώ είχε γίνει και κάποια συζήτηση για πιθανή (εξελικτικά) προέλευση μέσω διπλασιασμού για κάποιες οικογένειες [Volbeda και Hol, 1989].

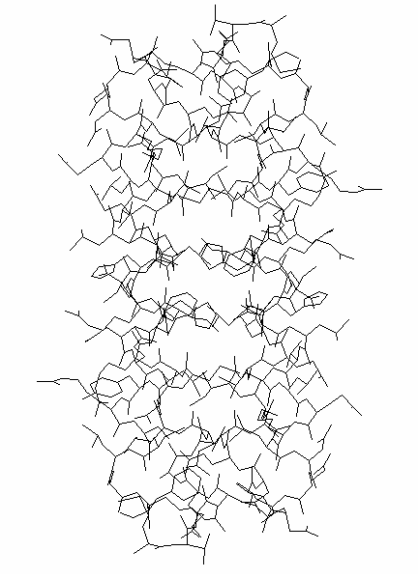
Στην πραγματικότητα, οι τέσσερις α-έλικες που το αποτελούν πακετάρονται μεταξύ τους υπό γωνία, γύρω από τον υδρόφοβο πυρήνα που σχηματίζουν, οδηγώντας σε μια συστροφή γύρω από τον κεντρικό άξονα του δεματιού (Εικόνα 2), είτε αριστερόστροφα (L στην εικόνα) ή δεξιόστροφα (R στην εικόνα).



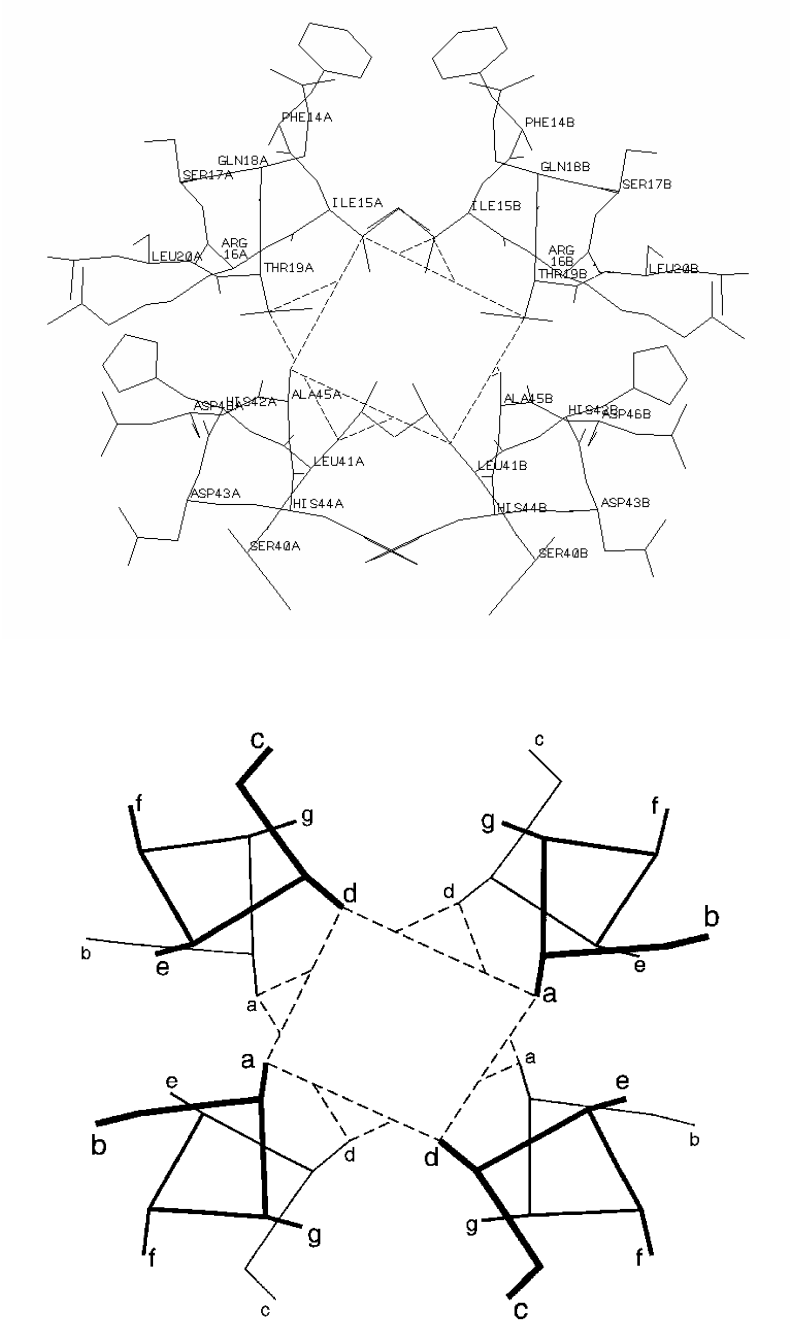
Εικ. 6. *Σχηματική αναπαράσταση των δύο εννοιών συστροφής του δεματιού.*
*(**L:**αριστερόστροφο, **R:**δεξιόστροφο.)*

Επειδή, όπως φάνηκε σε μια προκαταρκτική ανάλυση, οι δύο διαφορετικές αυτές συστροφές οδηγούν σε διαφορετικές τοπολογίες, αποφασίστηκε για την πρώτη αυτή φάση της ανάλυσης το δείγμα να περιοριστεί σε αριστερόστροφα μόρια, δεδομένου μάλιστα ότι τα δεξιόστροφα περιλαμβάνουν ορισμένα έντονα παραμορφωμένα.

Η τοπολογία του δεματιού είναι εξαιρετικά απλή. Οι τέσσερις έλικες που το αποτελούν είναι αμφιπαθικές (δηλαδή υδρόφιλες από τη μια πλευρά και υδρόφοβες από την άλλη), που σημαίνει ότι σε επίπεδο αλληλουχίας παρουσιάζουν μια εναλλαγή υδρόφιλων και υδρόφοβων αμινοξικών καταλοίπων. Και αυτή η εναλλαγή περιγράφεται σαν μια επανάληψη της μορφής $(abcdefg)_n$, όπου οι θέσεις a και d είναι υδρόφοβες (Εικόνα 7), οι b, c και f υδρόφιλες, με τις e και g στα όρια μεταξύ του υδρόφοβου πυρήνα και του εξωτερικού χώρου. Η διαφορά στα περιβάλλοντα των θέσεων είναι προφανής, και μένει να δει κανείς αν υπάρχουν πιο συγκεκριμένες προτιμήσεις πέρα από τη γενικά αναμενόμενη κατανομή υδρόφοβων-υδρόφιλων αμινοξικών τύπων. Η εσωτερική επανάληψη, δε, εκτός από την απλή τοπολογία στην οποία οδηγεί, αυξάνει τον αριθμό ανεξάρτητων παρατηρήσεων για κάθε θέση, βοηθώντας μια τέτοια στατιστική ανάλυση.



Εικ. 7α*. Πλαϊνή άποψη της πρωτεΐνης ROP, στην οποία φαίνεται μια χαρακτηριστική οργάνωση κατά στρώματα, "φέτες αλληλεπιδράσεων", δύο από τις οποίες (οι κεντρικές) δείχνονται σε μεγαλύτερη λεπτομέρεια στην Εικόνα 7β, καθώς θα μας απασχολήσουν σε επόμενα κεφάλαια.*



Εικ. 7β. *Τμήμα της πρωτεΐνης ROP σε δύο διαφορετικές απεικονίσεις (προκύπτουν στρίβοντας το κάτω μέρος της Εικόνας 7α κατά 90^{O} προς τον αναγνώστη): μια με όλα τα άτομα, και μια μόνο με τα α- και β- άτομα άνθρακα. Οι διακεκομμένες γραμμές ορίζουν επίπεδα κάθετα στον άξονα του δεματιού. Μάλιστα, για ορισμένα από τα μόρια που υιοθετούν το πρότυπο του δεματιού, οι περισσότερες από τις υδρόφοβες επαφές συμβαίνουν κοντά και*

παράλληλα στα επίπεδα αυτά. Το πλήθος των επαφών αποτελεί τη βάση της σταθερότητας του προτύπου.

Τι ακολουθεί

Η δυνατότητα, να καθρεφτίζεται η πρωτεϊνική δομή, με τη μορφή προτιμήσεων για αμινοξικούς τύπους, σε σύνολα θέσεων που ακολουθούν απλά πρότυπα κατά μήκος της αλληλουχίας, ορίζει αυτόματα έναν *στερεοχημικό κώδικα*, που αντιστοιχίζει δομές και αλληλουχίες. Βασικό χαρακτηριστικό του κώδικα αυτού αποτελεί ο *πλεονασμός (redundancy)*, με την έννοια ότι πολλές διαφορετικές (με οποιοδήποτε άλλο μέτρο ομοιότητας) αλληλουχίες διπλώνουν με τον ίδιο τρόπο, όπως διαπιστώνεται και από το πλήθος των "ανώδυνων" και χωρίς "καταστροφικά αποτελέσματα" μεταλλαγών που μπορεί να αντέξει οποιαδήποτε φυσική πρωτεΐνη και από την ύπαρξη οικογενειών πρωτεϊνών. Αυτό έχει δώσει τη βάση για τη μελέτη των σχέσεων δομής-αλληλουχίας από μια σκοπιά γνωστή ως *"αντίστροφο πρόβλημα"*, δηλαδή να βρεθούν αλληλουχίες συμβατές με μια δεδομένη δομή. Συνήθως, σ' αυτή την περίπτωση, προσδιορίζεται εμπειρικά (στατιστικά) η συμβατότητα κάθε θέσης μιας δομής, παρά ενός δομικού προτύπου, με τους 20 αμινοξικούς τύπους. Στη συνέχεια επιχειρείται σχεδιασμός αλληλουχιών συμβατών με τη δομή αυτή. Σε μια παραλλαγή, αν είναι γνωστό ότι μια δεδομένη αλληλουχία διπλώνει με τον συγκεκριμένο τρόπο, επιχειρείται η καλύτερη δυνατή αντιστοίχιση των καταλοίπων στις πιθανές θέσεις τους *(threading)*.

Η παρούσα μελέτη υπάγεται στη γενικότερη κατεύθυνση του αντίστροφου προβλήματος. Στο πρώτο μέρος, αναλύονται οι κατανομές των διαφόρων αμινοξικών τύπων στις τοπολογικά διακριτές θέσεις του δεματιού. Επιπλέον, διερευνώνται οι επιδράσεις των κατανομών αυτών στα τμήματα εκείνα που παραδοσιακά θεωρούνται "τυχαία δομή" και δείχνεται πως οι διαφορετικές κατανομές ταιριάζουν μεταξύ τους. Στο δεύτερο μέρος δίνονται διάφορες εφαρμογές των αποτελεσμάτων αυτών σε θέματα σχεδιασμού πρωτεϊνικών δομών, ενώ συζητούνται επιπτώσεις σε θέματα μοριακής εξέλιξης.

Αξίζει να σημειωθεί ότι η έννοια των τοπολογικά διακριτών θέσεων είναι ανεξάρτητη από συγκεκριμένα δομικά πρότυπα, αφού αναφέρεται στις σχέσεις της κάθε θέσης με τους γείτονές της στο χώρο, κάτι που ανοίγει τον δρόμο για πιθανή γενίκευση της προσέγγισης για οποιοδήποτε πρότυπο, και τη γενικότερη κατανόηση των σχέσεων δομής-αλληλουχίας.

Προσάρτημα: Οι πηγές σταθερότητας της δομής των πρωτεϊνών

Παρατίθενται τα βασικά χαρακτηριστικά αλληλεπιδράσεων σχετικών με τη δομή των πρωτεϊνών, για λόγους πληρότητας της Εισαγωγής, και κυρίως της συζήτησης σε προηγούμενη παράγραφο. Τα περισσότερα προέρχονται από το άρθρο του Dill [1990], ενώ η πλειοψηφία των ποσοστών για τους δεσμούς υδρογόνου απο το άρθρο των Baker και Hubbard [1984]. Άλλες λεπτομέρειες προέρχονται από το άρθρο ανασκόπησης των Richardson και Richardson [1989].

Δεσμοί υδρογόνου: Ο δεσμός R(OH)---(OH)R στο κενό, έχει ισχύ της τάξης των 6 $Kcal/mol_{Hbond}$. Σε μια πρωτεΐνη, όμως, και σε πειράματα μεταλλαξογένεσης, η σταθερότητα πέφτει κατά 0,5-1,5 $Kcal/mol_{Hbond}$ για κοινούς δεσμούς υδρογόνου που χάνονται (και ίσως περισσότερο από 4-5 $Kcal/mol_{Hbond}$ αν εμπλέκονται φορτισμένοι δότες/δέκτες). Αυτή η διαφορά οφείλεται, αφ'ενός, στο γεγονός ότι οι ίδιοι (και ίσως περισσότεροι) δεσμοί υδρογόνου θα μπορούσαν να γίνουν με το νερό (και ίσως με καλύτερη γεωμετρία, αφού τα μόρια του νερού έχουν μεγαλύτερη ελευθερία κινήσεως) και αφ'ετέρου, όταν χάνονται κάποιοι για την πρωτεΐνη, συχνά δημιουργούνται (και στη φυσική δομή πάντα) άλλοι μεταξύ πρωτεΐνης-νερού. Οι πιθανοί δότες/δέκτες, όμως, σε μια πρωτεΐνη είναι πολλοί, και όσοι περισσότεροι ικανοποιούνται στη διπλωμένη δομή, τόσο πιο ανταγωνιστική γίνεται η τελευταία έναντι της κατάστασης αποδιάταξης. Αυτό επέτρεψε και την πρόταση με θεωρητικά μέσα της α-έλικας και του β-πτυχωτού φύλλου σαν πιθανές δομές, αφού οι δεσμοί υδρογόνου, μαζί με τη στερεοχημική παρεμπόδιση, καθορίζουν το δίπλωμα στο β'-ταγές επίπεδο, καθώς δεν υπάρχουν πολλές στερεοδιατάξεις που να οδηγούν στο να ικανοποιηθούν όσο το δυνατό περισσότεροι. Μόνο περίπου 10% των πιθανών δοτών/δεκτών μιας πρωτεΐνης δεν σχηματίζει δεσμούς υδρογόνου παρατηρήσιμους στις κρυσταλλικές δομές, ενώ το 80% των δοτών/δεκτών της πρωτεΐνης που σχηματίζουν δεσμό υδρογόνου μέσα στην πρωτεΐνη εμπλέκουν την κύρια αλυσίδα. Ένα ποσοστό των ομάδων CO και NH της κύριας αλυσίδας (που ποικίλει στις διάφορες πρωτεΐνες μεταξύ 40%-70%) προσδένονται μεταξύ τους, με αξιοσημείωτο το γεγονός ότι η μεγαλύτερη τιμή αφορά τις πρωτεΐνες που αποτελούνται αποκλειστικά από α-έλικες. Από την άλλη, 45% των δοτών/δεκτών μιας πρωτεΐνης σχηματίζουν παρατηρήσιμους δεσμούς υδρογόνου με το νερό, πράγμα που κάνει άλλωστε τις πρωτεΐνες υδατοδιαλυτές. Τέλος, περίπου 10% των δεσμών υδρογόνου προς ή από τις ομάδες CO και NH της πρωτεΐνης εμπλέκουν πλευρικές αλυσίδες. Καθώς όμως στο συνολικό ισοζύγιο η διαφορά της ελεύθερης ενέργειας από δεσμούς υδρογόνου είναι σε βάρος της διπλωμένης δομής, σίγουρα δεν αποτελούν -ενεργειακά- κινητήριο δύναμη για να διπλώσει μια πρωτεΐνη, αν και -όπως αναφέρθηκε- μπορεί να οφείλεται, σε κάποιους

από αυτούς και σε κάποιες περιπτώσεις, μέρος της ειδικότητας των αλληλεπιδράσεων στη διπλωμένη δομή (όταν πχ βρίσκονται θαμμένοι σε υδρόφοβο περιβάλλον)· όχι όμως και του διπλώματος γενικότερα.

Άλλες ηλεκτροστατικές αλληλεπιδράσεις μέσα στο πρωτεϊνικό μόριο περιλαμβάνουν τα ζεύγη ιόντων, περιπτώσεις δηλαδή όπου δύο ομάδες με καθαρό φορτίο πλησιάζουν σε απόσταση μικρότερη από 4 □ (Δες Barlow D.J και Thornton J.M. (1983) *J.Mol.Biol.* **168**, 867- για μια πλήρη συζήτηση). Στο 37% των περιπτώσεων είναι μεταξύ ομωνύμων φορτίων. Ανάλογα με τους αμινοξικούς τύπους, 38% των Arg, 29% των His, 20% των Lys, και 16% των Asp συμμετέχουν σε ζεύγη ιότων. Αλλαγές στην αλληλουχία που προκαλούν απώλεια ζευγών ιόντων αποσταθεροποιούν την πρωτεΐνη κατά περίπου 3 $Kcal/mol_{\text{ζεύγους ιόντων}}$. Ακραίες τιμές pH οδηγούν σε μεγάλες τιμές καθαρού φορτίου και συνεπώς σε αποδιάταξη. Σημαντική διαφωνία υπάρχει σχετικά με την τιμή της φαινομένης διηλεκτρικής σταθεράς και τιμές που ποικίλουν από 2 εως 80 έχουν χρησιμοποιηθεί. Το να συμπεριλάβει κανείς ηλεκτροστατικά δεδομένα σε υπολογισμούς ενεργειών είναι από τα πράγματα που απαιτούν μεγάλη προσοχή. Άλλωστε, το 85% είναι στην επιφάνεια και όχι καλά συντηρημένα, που (μαζί με τη μη-καθολικότητα τους, γενικότερα) δείχνει ότι η σημασία τους για την δομή είναι μάλλον μικρή. Τέλος, μια άλλη πηγή ηλεκτροστατικού φορτίου στις πρωτεΐνες προκύπτει από το γεγονός ότι όλοι οι πεπτιδικοί δεσμοί στις α-έλικες έχουν τον ίδιο προσανατολισμό. Αυτό δημιουργεί ένα μακρο-δίπολο, σαν να υπάρχει θετικό φορτίο 0,5-0,75 μονάδες στο αμινοτελικό άκρο της έλικας και αντίστοιχο αρνητικό στο καρβοξυτελικό, είναι δε ικανό να προσδέσει ιόντα απουσία κοντινών φορτισμένων ομάδων, ή να μετατοπίσει την pK κοντινών φορτισμένων καταλοίπων (πχ της ιστιδίνης από 6.6 σε 7.8 στο καρβοξυτελικό άκρο). Οι όποιες αλληλεπιδράσεις με αυτό το δίπολο, όμως, όπως συμβαίνει και με τα περισσότερα ζεύγη ιόντων συμβαίνουν αφού διπλωθεί η πρωτεΐνη, και συνεπώς δεν κατευθύνουν το δίπλωμα.

Οι *δισουλφιδικοί δεσμοί* αποτελούν μια ακόμη πηγή σταθερότητας. Η ισχύς τους ποικίλει, και όταν είναι θαμμένοι στο εσωτερικό του μορίου είναι πιο σταθεροί. Καθώς δεν μπορούν να σχηματιστούν μεταξύ καταλοίπων της ίδιας α-έλικας, και μόνο υπό προϋποθέσεις στο ίδιο β-φύλλο, συνήθως ενώνουν τμήματα μακρινά στην αλληλουχία, και συνεπώς σταθεροποιούν τη δομή αφού διπλωθεί. Μερικές φορές συγκρατούν τμήματα της αλληλουχίας που αποτελούν ξεχωριστές αλυσίδες μετά από πρωτεολυτική διάσπαση (πχ ινσουλίνη). Σε πειράματα επαναδιάταξης, όταν μπορούν να σχηματισθούν δισουλφιδικοί δεσμοί από εναλλακτικά ζεύγη κυστεϊνών, συνήθως το δίπλωμα παγιδεύεται σε λάθος μονοπάτια, εκτός εαν δίνεται η δυνατότητα να σπάσουν οι λάθος δισουλφιδικοί δεσμοί και να δοκιμαστούν εκ νέου άλλοι. Αν ενώνουν χωριστές αλυσίδες

και σπάσουν, επέρχεται αποδιάταξη, και -απουσία του τμήματος που ένωνε τις αλυσίδες πριν την πρωτεόλυση- η επαναδιάταξη αποδεικνύεται αδύνατη. Σε πειράματα αποδιάταξης, όσο δεν διασπώνται, δεν επηρεάζουν την κινητική του πειράματος. Αυτά δείχνουν ότι, ακόμη κι όπου υπάρχουν, δεν κατευθύνουν το δίπλωμα (μάλλον το μπερδεύουν!)· επίσης, λαμβάνοντας υπ'όψη την ευκολία με την οποία σπάνε σε αναγωγικά περιβάλλοντα, και τη μη-καθολικότητα τους, ότι αν συνεισφέρουν ενεργειακά στο δίπλωμα η συνεισφορά αυτή είναι μάλλον ελάχιστη.

Το περιβάλλον στο οποίο διπλώνει μια πρωτεΐνη (πολικότητα του διαλύτη, οξύτητα, άλλες διαλυμένες μικρο- και μακρομοριακές ουσίες) μπορεί να παίζει σημαντικό ρόλο στο βαθμό στον οποίο σταθεροποιούν οι παραπάνω δυνάμεις μια δομή, και δεν είναι πάντα εύκολο να ληφθεί υπ'όψη. Σαν ακραίο παράδειγμα συνήθως αναφέρεται η περίπτωση των μεμβρανικών πρωτεϊνών, αλλά και οι σφαιρικές υδατοδιαλυτές πρωτεΐνες συχνά αντιμετωπίζουν ένα ακραίο περιβάλλον, όπως η οξύτητα του στομαχιού, το αναγωγικό περιβάλλον στο εσωτερικό μιας E. coli κ.ά. Στην απλούστερη περίπτωση, η οξύτητα του διαλύματος καθορίζει το ολικό φορτίο μιας πρωτεΐνης, ενώ η ιοντική ισχύς του καθορίζει το πόσο ισχυρά αλληλεπιδρούν τα φορτία μεταξύ τους. Το είδος του διαλύτη και οι ιδιότητές του είναι που καθορίζουν την κατανομή των πλευρικών αλυσίδων στο εκάστοτε "εσωτερικό" ή "εξωτερικό" του μορίου, και οι υδρόφοβες αλληλεπιδράσεις, η κύρια συνιστώσα στη σταθερότητα των σφαιρικών υδατοδιαλυτών πρωτεϊνών, αποτελούν υποσύνολο της περίπτωσης αυτής. Ας σημειωθεί ότι οι έννοιες "εσωτερικό/ εξωτερικό του μορίου" δεν είναι πάντα τόσο καθαρές. (Για παράδειγμα σε ένα κανάλι μεταφοράς ιόντων κάθετα σε μια μεμβράνη ποιο πρέπει να θεωρείται εσωτερικό και ποιο εξωτερικό;) Όμως, ακόμα και σε ακραίες περιπτώσεις, οι τύποι διπλώματος που παρατηρούνται είναι (μέχρι τώρα τουλάχιστο) υποσύνολο των ήδη γνωστών και όχι κάτι δραστικά καινούριο. Από την άλλη, ακόμη και για την εκτενώς μελετημένη περίπτωση του ύδατος, σε σχέση με τις σφαιρικές υδατοδιαλυτές πρωτεΐνες, τα πράγματα δεν είναι ξεκάθαρα. Η κατανομή που "επάγει" με βάση την υδροφοβικότητα του κάθε αμινοξικού τύπου δεν οδηγεί σε πλήρη διαχωρισμό, αφού κάποια υδρόφοβα μένουν στην επιφάνεια, κάποια υδρόφιλα θάβονται (πχ. σε ενεργά κέντρα) ενώ κάποια άλλα (υδρόφιλα) έχουν αρκετά μακριές αλυσίδες (πχ η λυσίνη) ώστε το υπόλοιπο τμήμα τους να είναι στο εσωτερικό και το φορτίο στην επιφάνεια. Η περιοδικότητα που παρουσιάζει (ανά 3.6 κατάλοιπα στην α-έλικα και ανά 2 στα β-πτυχωτά φύλλα) και που συνήθως χρησιμοποιείται για πρόβλεψη δομής, μπορεί εύκολα να βγει εκτός φάσης από μια β-διόγκωση, τις ακραίες στροφές των α-ελίκων ή εισδοχές ενός καταλοίπου σε μια α-έλικα. Επίσης ας σημειωθεί ότι, σε θεωρητικούς ενεργειακούς υπολογισμούς, το περιβάλλον με τη γενικότερη έννοια, είναι αυτό που συνήθως λαμβάνεται λιγότερο (εως καθόλου) υπ'όψη.

"Κανείς δεν μπορούσε να πιστέψει, ότι έτρεχα έτσι, χωρίς λόγο."

(Σε ελεύθερη απόδοση από την ταινία "Forrest Gump")

Μέρος Α / Κεφάλαιο Ι:

Κατανομές κατά την ακτινική έννοια

Εισαγωγή

Στη Γενική Εισαγωγή, με αφετηρία το γεγονός ότι οι πρωτεΐνες οργανώνονται με βάση λίγα (ίσως 100) τριτοταγή δομικά πρότυπα, διατυπώθηκε μια ιδέα, βασισμένη σε απλή διαισθητική αντίληψη: εαν υπάρχουν κοινά χαρακτηριστικά μεταξύ των προτιμήσεων των τοπολογικά διακριτών θέσεων ενός δομικού προτύπου για συγκεκριμένους αμινοξικούς τύπους, σε πρωτεΐνες που δεν παρουσιάζουν άλλη ομοιότητα εκτός του ότι υιοθετούν το ίδιο αυτό δομικό πρότυπο σαν μέρος (ή σύνολο) της τριτοταγούς δομής τους, τότε -λογικά- θα πρέπει να αντιπροσωπεύουν γενικά χαρακτηριστικά που πρέπει να ικανοποιεί μια πρωτεϊνική αλληλουχία, ώστε να είναι συμβατή με το πρότυπο αυτό, ανεξάρτητα από λειτουργία και προέλευση.

Στο παρόν "Μέρος Α" αναζητείται η ύπαρξη τέτοιων προτιμήσεων στο απλό δομικό πρότυπο, που παρουσιάστηκε στη Γενική Εισαγωγή, το αντιπαράλληλο δεμάτι από τέσσερις α-έλικες στην αριστερόστροφη εκδοχή του (και που για συντομία θα αναφέρεται απλά σαν "δεμάτι") και υπολογίζεται κατ' αρχήν το μέγεθός τους. Στο Κεφάλαιο Ι, η α-έλικα εξετάζεται ακτινικά (δηλαδή βλέποντάς την από την μια άκρη, οπότε οι επτά θέσεις του προβάλλουν ολόγυρα σαν ακτίνες). Η εργασία αυτή έγινε την περίοδο Δεκέμβριος 1989-Καλοκαίρι 1990 και είναι δημοσιευμένη [Paliakasis και Kokkinidis, 1991 και 1992]. Στο Κεφάλαιο ΙΙ εξετάζεται αξονικά (δηλαδή βλέποντας τον άξονα της α-έλικας "από το πλάι" και εξετάζοντας τι γίνεται κατά μήκος του), και επιπλέον εξετάζονται και τα συνδετικά μεταξύ των α-ελίκων τμήματα.

Η πιθανότητα, τέτοιες προτιμήσεις, αν και χαμηλές για να επιτρέψουν συστοίχιση αλληλουχιών, να είναι ικανές να επιτρέψουν την ασφαλή κατάταξη μιας αλληλουχίας σε ένα δομικό πρότυπο, όπως επίσης εαν είναι ειδικές για το συγκεκριμένο πρότυπο ή διαφορετικά πρότυπα μπορούν να έχουν παρόμοιες προτιμήσεις, διερευνάται σε επόμενα κεφάλαια.

Διαδικασία

Επιλογή δείγματος: Το πρώτο βήμα, μετά την επιλογή του δομικού προτύπου για το οποίο θα γίνει η ανάλυση, είναι η δημιουργία μιας μικρής βάσης πληροφοριών από μόρια όπου παρατηρείται το πρότυπο αυτό σαν μέρος ή σύνολο της τριτοταγούς δομής τους. Το δείγμα που χρησιμοποιήθηκε δίνεται στον Πίνακα 1.

Πίνακας 1. *Τα τμήματα των αντιπροσώπων (με γνωστή δομή) των επτά οικογενειών που χρησιμοποιήθηκαν στην παρούσα ανάλυση. (Σύνολο: 523 θέσεις.)*

Οικογένεια	Κωδ. PDB	α-έλικα 1	α-έλικα 2	α-έλικα 3	α-έλικα 4
ROP	1ROP	3K-29L	31A-56F		
Φερριτίνη		14Q-43R	48L-76R	96G-124K	127P-158M
Αιμερυθρίνη	1HMQ	19T-38A	41A-64A	70Y-86T	91V-109Y
(Μυοαιμερυθρίνη)	(2MHR)	19E-38D	41A-64A	70V-86G	96V-114Y
Κυτόχρωμα c’	2CCY	5P-31A	39D-54A	79S-102A	104P-125F
Κυτόχρωμα β_{562}	156B	2D-18E	24K-46N	62K-84E	87V-108K
Λυσοζύμη φάγου T4	3LZM	93A-106M	115T-122Q	126W-133L	143P-154R
Πρωτ. καλύμματος ιού μωσαϊκής καπνού	2TMV	21I-30A	39Q-50E	74A-87F	117A-134R

Οι ατομικές συντεταγμένες των πειραματικά (κυρίως κρυσταλλογραφικά) προσδιορισμένων δομών των μορίων αυτών προήλθαν από την Protein Data Bank [Bernstein *et al*, 1977], εκτός από την περίπτωση της φερριτίνης, όπου οι συντεταγμένες *ήταν*[15] ευγενική προσφορά της Prof. P. Harrison.

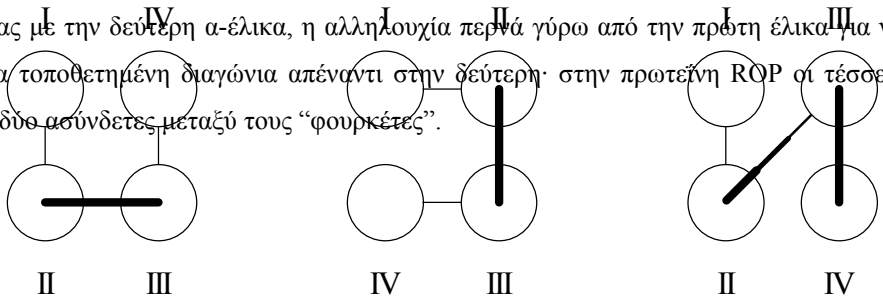
Όσον αφορά τα άκρα των α-ελίκων, όπως συνήθως είναι η περίπτωση σε αναλύσεις του είδους, μερικές φορές υπήρχε αμφιβολία αν ένα αμινοξικό κατάλοιπο αποτελούσε μέρος της έλικας ή όχι. Αποφασίστηκε ότι ένα κατάλοιπο θα συμπεριλαμβάνοταν στην ανάλυση, αν μπορούσε να πιστοποιηθεί από τις αποστάσεις της πλευρικής του ομάδας από γειτονικές της στο χώρο και το πακετάρισμά της, ότι εκτελεί ρόλο κάποιας από τις θέσεις a-g, με τον α-άνθρακα “κοντά” (οπτικά) στην τροχιά της α-έλικας. Από μια τέτοια θέση, συνήθως συμμετέχει σε έναν υδρογονοδεσμό τύπου έλικας (α- ή 3_{10}), χωρίς αυτό να αποτελεί απαίτηση.

Τα συνδετικά τμήματα εξαιρέθηκαν από αυτή την αρχική φάση της ανάλυσης, εξαιτίας της ποικιλότητας που παρουσιάζουν, αλλά αναλύθηκαν στη συνέχεια και καλύπτονται σε επόμενο κεφάλαιο. Τα μόρια αυτά καλύπτουν διάφορες μορφές *συνδεσμολογίας*, δηλαδή σχετικής θέσης των συνδετικών τμημάτων μεταξύ των α-ελίκων (Εικόνα 1).

Στα περισσότερα τέτοια τμήματα, προχωρώντας κανείς στην αλληλουχία πέρα από το καρβοξυτελικό άκρο κάθε έλικας, βρίσκεται στο αμινοτελικό άκρο της γειτονικής του· στη φερριτίνη πάλι,

[15] Στο μεταξύ, έχει κατατεθεί (στην PDB) η δομή της βαριάς αλυσίδας της φερριτίνης του ανθρώπου με κωδικό 1FHA. Στο Μέρος Β υπάρχει κατάλογος με (πειραματικά προσδιορισμένες) δομές δεματιών, μετά 5 έτη, το 1995.

τερματίζοντας με την δεύτερη α-έλικα, η αλληλουχία περνά γύρω από την πρώτη έλικα για να συνεχίσει σε μια έλικα τοποθετημένη διαγώνια απέναντι στην δεύτερη· στην πρωτεΐνη ROP οι τέσσερις α-έλικες ανήκουν σε δύο ασύνδετες μεταξύ τους "φουρκέτες".



Εικ. 1. *Σχηματική αναπαράσταση πιθανών συνδεσμολογιών σε δεμάτια. (Οι έντονες γραμμές αντιστοιχούν σε συνδέσεις στο εμπρός μέρος της σελίδας, και οι λεπτές στο πίσω.) Η πρώτη είναι αριστερόστροφη συνεχής (παράδειγμα: λυσοζύμη του φάγου T4), ενώ η δεύτερη δεξιόστροφη συνεχής (παράδειγμα: αιμερυθρίνη)· η τρίτη -αριστερόστροφη- είναι παράδειγμα μη συνεχούς σύνδεσης, και απαντά στην φερριτίνη· το διαγώνιο συνδετικό τμήμα στην πραγματικότητα περνά γύρω από την πρώτη α-έλικα, ενώ στην πρωτεΐνη ROP δεν υπάρχει.*

Αντίθετα, δεν έγινε καμία προσπάθεια εξαίρεσης "ειδικών" περιπτώσεων (πχ ενεργά κέντρα), εφ'όσον ανήκαν στα τμήματα που αναφέρονται στον Πίνακα 1, με το σκεπτικό ότι η εκεί παρουσία τους είναι πρώτα συμβατή με τις στερεοδιαταξικές και θερμοδυναμικές ανάγκες της θέσης όπου βρέθηκαν. Η πιθανότητα αυτές οι τελυταίες να απέχουν από τους συνήθεις μέσους όρους του δεματιού, απλά εισάγει στα αποτελέσματα (αν και με τη μορφή "θορύβου") ένα μέτρο της ποικιλότητας που μπορεί να παρουσιάσει μια θέση και -πιο συνολικά- τις αποκλίσεις που μπορεί να αναμένει κανείς.

Τέλος, οι θέσεις a-g προσδιορίστηκαν με άμεση παρατήρηση σε μια οθόνη γραφικών, αφού σε αυτή την πρώτη φάση η έννοια των τοπολογικά διακριτών θέσεων είχε οριστεί περισσότερο διαισθητικά (κάτι που σύντομα άλλαξε -περισσότερα στο Κεφ.Α.ΙΙ), και δεν υπήρχαν αντικειμενικά κριτήρια που να κατατάσσουν μια θέση στο είδος της με βάση κάποια χαρακτηριστικά της.

Επέκταση δείγματος: Εκτός των μορίων που η δομή τους είχε προσδιοριστεί πειραματικά, για τέσσερις από τις επτά οικογένειες, υπήρχαν αλληλουχίες, σε βάσεις δεδομένων, ομόλογες σε βαθμό που να επιτρέπεται η μεταφορά των επτά θέσεων και σε αυτές, με αλληλουχιακές συστοιχίσεις (Πίνακας 2)· η σημασία τους συζητιέται στο τέλος του κεφαλαίου.

Η αναζήτηση των ομόλογων αλληλουχιών έγινε με το πρόγραμμα FASTA [Pearson και Lipman 1988] στην έκδοση 27 της βάσης δεδομένων NBRF [George *et al*, 1988] και στην έκδοση 16 της βάσης δεδομένων SwissProt [Bairoch και Boeckman, 1991], οι δε πολλαπλές συστοιχίσεις με το πρόγραμμα CLUSTAL [Higgins και Sharp, 1988]. Και στις τέσσερις περιπτώσεις πρόκειται στην πραγματικότητα για τις ίδιες (τέσσερις) πρωτεΐνες από διαφορετικούς οργανισμούς (ή στελέχη).

Επεξεργασία: Στη συνέχεια, μετρήθηκε ο αριθμός των περιπτώσεων (*απόλυτη συχνότητα*) που απαντά καθένα από τα είκοσι αμινοξικά κατάλοιπα σε καθεμιά από τις επτά θέσεις, για κάθε οικογένεια χωριστά. Για κάθε οικογένεια έγιναν δύο κανονικοποιήσεις: Η μία ήταν διαίρεση της απόλυτης συχνότητας κάθε τύπου αμινοξέος στην κάθε θέση με το άθροισμα των απολύτων συχνοτήτων των είκοσι τύπων στην θέση αυτή, που δίνει τη *σχετική σύσταση* της θέσης στον τύπο αυτό. Η άλλη ήταν διαίρεση της απόλυτης συχνότητας κάθε τύπου αμινοξέος στην κάθε θέση με το άθροισμα των απολύτων συχνοτήτων του τύπου αυτού στις επτά θέσεις, που δίνει την σχετική συχνότητα (*προτίμηση*) του τύπου αυτού στη θέση αυτή.

Πιο συγκεκριμένα, για κάθε οικογένεια ***f*** δημιουργήθηκε ένας πίνακας $\mathbf{R}_f(i,k)$ (όπου ***i***=1-20 οι 20 τύποι αμινοξικών καταλοίπων και ***k***=1-7 οι 7 τύποι θέσεων), τα στοιχεία του οποίου δείχνουν πόσες φορές απαντά το κατάλοιπο ***i*** στη θέση ***k***. Στη συνέχεια, και για κάθε οικογένεια ***f*** υπολογίστηκαν δύο πίνακες: σχετικής σύστασης θέσεων $\mathbf{O}_f(i,k)$ (από τον αρχικό όρο *occurrence*) και προτιμήσεων $\mathbf{P}_f(i,k)$ (από τον αρχικό όρο *preference*). Τα στοιχεία των πινάκων αυτών δίνονται από τους τύπους:

$$\mathbf{O}_f(i,k) = \mathbf{R}_f(i,k) \,/\, \sum_{i=1\text{-}20} \mathbf{R}_f(i,k) \quad \text{...και...} \quad \mathbf{P}_f(i,k) = \mathbf{R}_f(i,k) \,/\, \sum_{k=1\text{-}7} \mathbf{R}_f(i,k)$$

Οι μέσοι όροι για τις δύο σειρές πινάκων υπολογίστηκαν από τους τύπους:

$$\mathbf{O}(i,k) = \sum_{f=1\text{-}7} \mathbf{O}_f(i,k) \,/\, 7 \quad \text{...και...} \quad \mathbf{P}(i,k) = \sum_{f=1\text{-}7} \mathbf{P}_f(i,k) \,/\, 7$$

Οι υπολογισμοί έγιναν με απλά προγράμματα, που παρατίθενται στο τέλος του κεφαλαίου.

Πίνακας 2. *Οι τέσσερις οικογένειες για τις οποίες υπήρχαν σε βάσεις δεδομένων αλληλουχίες ομόλογες προς την πειραματικά προσδιορισμένη δομή.*

Οικογένεια	Είδος	Κωδικός SwissProt	Κωδ. NBRF
Φερριτίνη			
	Homo sapiens (heavy chain)	FRIHHUMAN	FRHUH
	Rana catesbeina (heavy chain)	FRI1RANCA	FRFGL
	Homo sapiens (light chain)	FRILHUMAN	FRHUL
	Rattus norvegicus (light chain)	FRILRAT	FRRTL
	Equus caballus (light chain)	FRILHORSE	FRHOL
Αιμερυθρίνη			
[1HMQ]	Themiste dyscritum	HEMTTHEDY	HRTHBD
[1HRB]	Phascolopsis gouldii	HEMTPHAGO	HRGG
	Themiste zostericola	HEMTTHEZO	HRTH
Μυοαιμερυθρίνη			
[2MHR]	Themiste zostericola	HEMMTHEZO	HRTHM
Κυτόχρωμα c’			
[2CCY]	Rhospirillum molischianum	CYCPRHOMO	CCQFCM
	Rhodospirillum fulvum	CYCPRHOFU	CCQFCF
	Alcaligenes sp. (NCIB 11015)	CYCPALCSP	CCALC
	Rhodocyclus gelatinosus (ή Rhodopseudomonas gelatinosa)	CYCPRHOGE	CCRFCG
	Paracoccus sp. (ATCC 12084)	CYCPPARSP	CCPCC8
	Rhodobacter sphaeroides	CYCPRHOSH	CCRFCS
	Rhodospirillum rubrum	CYCPRHORU	CCQFCR
	Rhodospirillum photometricum	CYCPRHOPH	CCQFCP
	Rhodopseudomonas sp	CYCPRHOSP	CCRFCX
	Rhodobacter capsulatus (ή Rhodopseudomonas capsulata)	CYCPRHOCA	CCRFPP
TMV coat protein			
[2TMV]	TMV vulgare	COATTMV	VCTMVU
	TMV Kokubu	COATTMVO	
	TMV OM	COATTMVOM	
	TMV 06	COATTMV06	
	TMV ER	COATTMVER	
	TMV dahlemense	COATTMVDA	VCTMDA
	TMV tomato	COATTMVTO	
	TMV U2	COATTMGMV	VCTMU2
	TMV ORS	COATTMVOR	VCTMOR
	TMV HR	COATTMVHR	VCTMHR
	TMV cowpea	COATTMVCO	VCTMCP

Αποτελέσματα

Οι πίνακες σχετικής σύστασης και προτίμησης που παρατίθενται (Πίνακες 3,4) αποτελούν μέσους όρους από τις επτά οικογένειες[16]. Στους πίνακες αυτούς (και στις αντίστοιχες εικόνες), τα 20 αμινοξικά κατάλοιπα έχουν ταξινομηθεί κατά σειρά υδροφοβικότητας σύμφωνα με την κλίμακα των Lesser και συνεργατών [1987] ενώ οι τοπολογικά διακριτές θέσεις έχουν τοποθετηθεί ως εξής: πρώτα οι περισσότερο κρυμμένες στο εσωτερικό του προτύπου (a και d), μετά εκείνες που βρίσκονται στα όρια υδρόφοβου εσωτερικού και υδρόφιλου εξωτερικού (e και g), και τελευταίες οι εκτεθειμένες στον εξωτερικό χώρο (b, c και f).

> ***Σημαντική τεχνική σημείωση:*** παρά την έλλειψη ομοιότητας (σε επίπεδο συστοίχισης αλληλουχιών), μεταξύ των οικογενειών που χρησιμοποιήθηκαν, οι πίνακες των σχετικών συστάσεων και των προτιμήσεων είναι αρκετά ομοιόμορφοι μεταξύ των επτά οικογενειών: τα επικρατή χαρακτηριστικά που περιγράφονται στη συνέχεια, παρατηρούνται και στις επτά.

Ξεκινώντας από την προτίμηση κάθε αμινοξικού καταλοίπου, η γενική εικόνα είναι ότι ακολουθείται το αναμενόμενο πρότυπο με τα υδρόφοβα κατάλοιπα "μέσα" και τα υδρόφιλα "έξω" (Πίνακας 3 και Εικόνα 2). Με μια προσεκτικότερη ματιά όμως, διαπιστώνει κανείς ότι, για πολλά κατάλοιπα (Lys, Asp, Glu, Gly, Met, Leu, Trp και Cys), υπάρχει μια θέση για την οποία δείχνουν ισχυρή προτίμηση (30-40%), αν και ορισμένα δείχνουν την ίδια περίπου κατανομή και στις επτά θέσεις (Ala, Asn). Οι ασυμμετρίες στην κατανομή μεταξύ θέσεων a και d, όπως επίσης μεταξύ e και g, και τέλος μεταξύ b και c, ενδέχεται να αντανακλούν ασυμμετρίες στην τοπολογία των θέσεων αυτών και τη στερεομετρία του περιβάλλοντός τους (Εικόνα 7/Εισαγωγή). Για την θέση f τέλος, την πιο εκτεθειμένη και με τους λιγότερους περιορισμούς, οι προτιμήσεις είναι στις ισχυρότερες των περιπτώσεων μάλλον μέτριες.

Συνεχίζοντας με τη σχετική σύσταση κάθε θέσης, η εικόνα αλλάζει: για κάθε θέση, το 50% των περιπτώσεων καταλαμβάνεται από όχι περισσότερους από 5 αμινοξικά τύπους (Πίνακας 4 και Εικόνα 3). Ακραία γίνονται τα πράγματα στο εσωτερικό (στον υδρόφοβο πυρήνα), όπου περισσότερο από το 40% των θέσεων d καταλαμβάνεται από δύο μόνο τύπους καταλοίπων: Ala και Leu. Ο μικρός αυτός αριθμός κυρίαρχων αμινοξικών τύπων στο εσωτερικό (ενώ στον εξωτερικό χώρο υπάρχει μια ισορροπία που περιλαμβάνει πολλά -υδρόφιλα- αμινοξικά κατάλοιπα), είναι -μάλλον- ενδεικτικός των περιορισμών που τίθενται λόγω τοπολογίας και που -όπως αναμένεται- είναι ισχυρότεροι στο εσωτερικό.

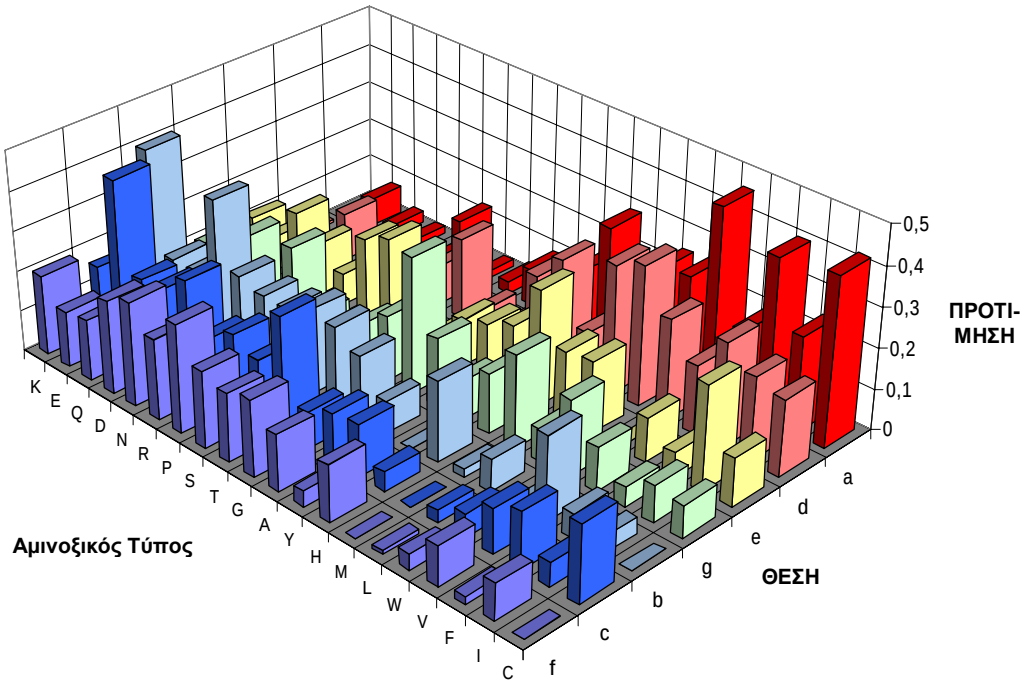
Πίνακας 3. *Η μέση προτίμηση των είκοσι αμινοξικών τύπων για καθεμία από τις επτά τοπολογικά διακριτές θέσεις a-g. Τιμές στην κλίμακα 0-1*

a	d	e	g	b	c	f

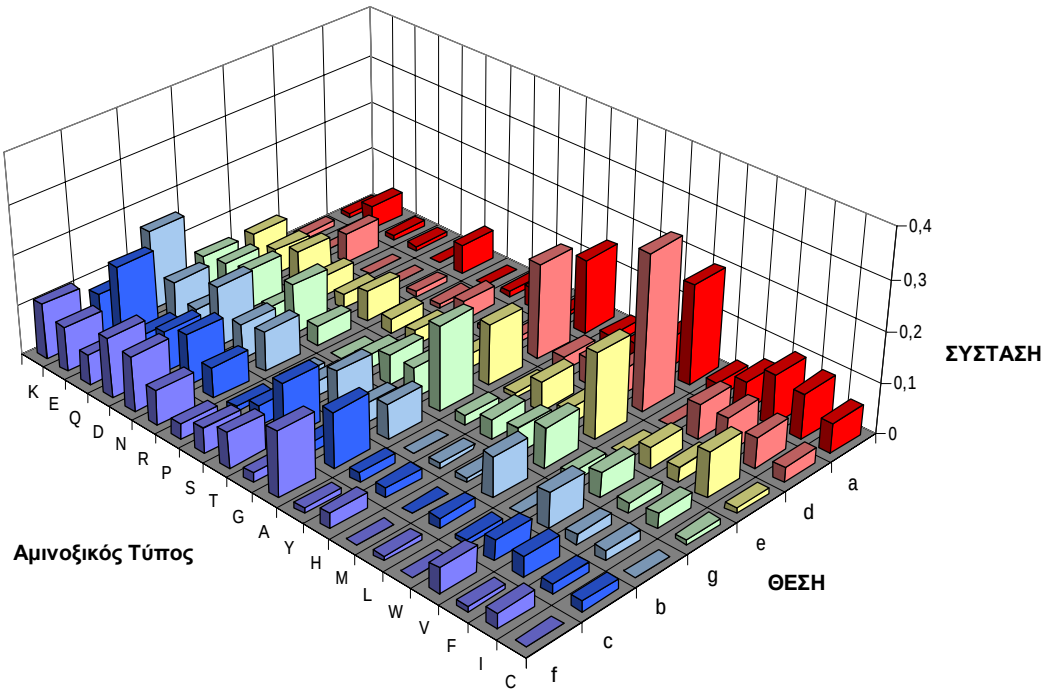
[16] Ειδικά στην περίπτωση των αιμερυθρινών, οι 5 αιμερυθρίνες θεωρήθηκαν σαν μια υπο-οικογένεια, η μυο-αιμερυθρίνη σαν μια δεύτερη, και εξήχθη μέσος όρος από τις δύο, ώστε να μην υπαρξει υπερ-εκπροσώπηση σε βάρος της μυοαιμερυθρίνης.

K	0.01	0.01	0.11	0.10	0.40	0.16	0.20
E	0.07	0.01	0.09	0.13	0.15	0.41	0.14
Q	0.03	0.12	0.18	0.18	0.15	0.19	0.15
D	0.02	0.01	0.13	0.05	0.36	0.20	0.23
N	0.00	0.00	0.08	0.21	0.20	0.25	0.26
R	0.13	0.08	0.20	0.07	0.18	0.14	0.20
P	0.02	0.10	0.23	0.01	0.18	0.18	0.27
S	0.03	0.20	0.11	0.11	0.21	0.15	0.19
T	0.09	0.05	0.05	0.16	0.20	0.29	0.17
G	0.06	0.04	0.12	0.34	0.16	0.09	0.19
A	0.15	0.20	0.14	0.17	0.08	0.12	0.14
Y	0.29	0.27	0.17	0.10	0.00	0.13	0.04
H	0.06	0.12	0.29	0.15	0.20	0.05	0.14
M	0.25	0.32	0.17	0.23	0.02	0.00	0.00
L	0.26	0.35	0.18	0.08	0.08	0.03	0.01
W	0.46	0.25	0.00	0.19	0.00	0.06	0.04
V	0.19	0.17	0.11	0.11	0.21	0.11	0.10
F	0.40	0.26	0.06	0.05	0.06	0.14	0.02
I	0.24	0.21	0.26	0.09	0.04	0.06	0.09
C	0.42	0.19	0.12	0.08	0.00	0.19	0.00

Ας σημειωθεί, ότι το 14% περίπου των θέσεων a και d καταλαμβάνονται από υδρόφιλα αμινοξικά κατάλοιπα, και αντίστοιχα το 15-20% των θέσεων b, c και f από υδρόφοβα. Ειδικά για το πρώτο, τα υδρόφιλα αυτά κατάλοιπα είτε βρίσκονται στην πρώτη ή την τελευταία στροφή της α-έλικας και άρα είναι εκτεθειμένα στον διαλύτη, είτε συνεισφέρουν σε ενεργά κέντρα, σημεία πρόσδεσης συμπαραγόντων ή/και προσθετικών ομάδων. Κατάλοιπα Phe και Trp σε θέσεις a και d (όπου φαινομενικά δεν υπάρχει χώρος για τις ογκώδεις και άκαμπτες πλευρικές τους αλυσίδες) μπορεί να σημαίνουν αυξημένη απόσταση μεταξύ δύο α-ελίκων, όπως στη φερριτίνη, ενώ σε κάποιες περιπτώσεις βρίσκονται σε τελικές στροφές και εκτελούν χρέη ενός είδους "καπακιού" που απομονώνει τον υδρόφοβο πυρήνα από τον εξωτερικό χώρο (Εικόνα 4).



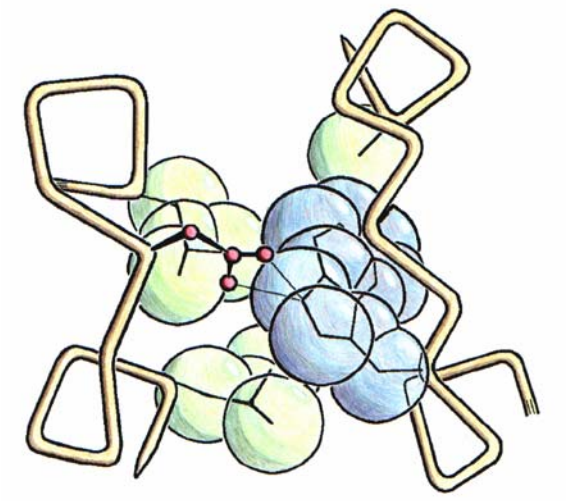
Εικ. 2. *Γραφική παράσταση της μέσης προτίμησης των είκοσι αμινοξικών τύπων για καθεμία από τις επτά τοπολογικά διακριτές θέσεις a-g.*



Εικ. 3. *Γραφική παράσταση της μέσης σύστασης σε αμινοξικούς τύπους για καθεμία από τις επτά τοπολογικά διακριτές θέσεις a-g.*

Πίνακας 4. *Η μέση σχετική σύσταση σε αμινοξέα για καθεμία από τις επτά τοπολογικά διακριτές θέσεις a-g. Τιμές στην κλίμακα 0-1*

	a	d	e	g	b	c	f
K	0.01	0.01	0.06	0.06	0.16	0.08	0.11
E	0.05	0.00	0.05	0.07	0.08	0.16	0.09
Q	0.01	0.06	0.07	0.08	0.05	0.05	0.06
D	0.01	0.00	0.04	0.03	0.12	0.08	0.12
N	0.00	0.00	0.03	0.10	0.07	0.10	0.11
R	0.06	0.01	0.06	0.04	0.08	0.06	0.07
P	0.00	0.01	0.03	0.00	0.01	0.01	0.03
S	0.01	0.04	0.03	0.03	0.03	0.04	0.05
T	0.04	0.01	0.04	0.06	0.08	0.11	0.07
G	0.01	0.01	0.03	0.06	0.04	0.02	0.02
A	0.15	0.19	0.12	0.17	0.07	0.11	0.13
Y	0.02	0.03	0.01	0.02	0.00	0.02	0.01
H	0.03	0.05	0.06	0.04	0.01	0.02	0.03
M	0.06	0.05	0.03	0.05	0.01	0.00	0.00
L	0.20	0.31	0.17	0.08	0.08	0.02	0.01
W	0.03	0.00	0.00	0.02	0.00	0.01	0.00
V	0.06	0.07	0.04	0.05	0.07	0.04	0.05
F	0.10	0.07	0.03	0.02	0.02	0.04	0.01
I	0.09	0.06	0.09	0.03	0.02	0.02	0.03
C	0.06	0.03	0.01	0.01	0.00	0.02	0.00



Εικ. 4. *Σχηματική αναπαράσταση της τοποθέτησης ενός αρωματικού αμινοξικού καταλοίπου σε ρόλο καπακιού στο άκρο ενός δεματιού. Η τρυπτοφάνη (μπλε) γεμίζει τον χώρο ανάμεσα στο υπερκείμενο συνδετικό τμήμα (καφέ) και τα υποκείμενα υδρόφοβα κατάλοιπα (πράσινα). Προφανώς δεν θα υπήρχε αρκετός χώρος να πακεταριστεί, αν ήταν στο ίδιο επίπεδο με αυτά. Οι υπόλοιπες πλευρικές αλυσίδες έχουν παραλειφθεί για λόγους ευκρίνειας. Το ασπαρτικό οξύ (κόκκινο) και οι δεσμοί υδρογόνου που σχηματίζει συζητούνται σε επόμενο κεφάλαιο. (Μόριο: αιμερυθρίνη).*

Τέλος, στον Πίνακα 5 η μέση σχετική σύσταση του δεματιού σε αμινοξικούς τύπους αντιπαραβάλλεται με εκείνη των πρωτεϊνών συνολικά και των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικες [Nakashima et al, 1986]. Γενικά, και με εξαίρεση λίγους αμινοξικούς τύπους[17], η σχετική σύσταση των επτά οικογενειών σε αμινοξικά κατάλοιπα δεν φαίνεται να διαφέρει σημαντικά, από εκείνη των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικες. Αντίθετα, όπως μπορεί να φανεί με μια σύντομη αναφορά στον Πίνακα 4, η σχετική σύσταση καθεμίας από τις επτά θέσεις διαφέρει από εκείνη των πρωτεϊνών συνολικά πολύ περισσότερο από ότι διαφέρει από την τελευταία η μέση σχετική σύσταση του δεματιού ή των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικες.

Πίνακας 5. *Μέση %σύσταση σε αμινοξικούς τύπους των πρωτεϊνών γενικά, των πρωτεϊνών που αποτελούνται μόνο από α-έλικες, και του δεματιού. Η μέση σύσταση του δεματιού, για κάθε αμινοξικό τύπο, υπολογίστηκε σαν μέσος όρος της σύστασης των επτά θέσεων για τον τύπο αυτό.*

	Πρωτεΐνες συνολικά	Πρωτεΐνες αποκλ. από α-έλικες	Δεμάτι
K	6.78	10.10	7.0
E	6.39	6.52	7.2
Q	3.91	3.33	5.1
D	5.72	6.52	5.9
N	4.39	3.79	5.7
R	4.81	2.79	5.8
P	4.49	3.81	1.3
S	6.56	5.44	3.2
T	5.84	4.91	5.8
G	7.82	7.66	2.7
A	8.74	11.63	13.4
Y	3.33	2.55	1.7
H	2.15	2.79	3.1
M	2.08	2.42	3.0
L	8.20	8.89	12.2
W	1.17	1.17	0.7
V	7.01	6.02	5.4
F	3.87	4.22	4.2
I	5.15	3.72	4.6
C	1.62	1.71	1.9

Συμπεράσματα-Συζήτηση

[17]Το δεμάτι παρουσιάζει μειωμένη περιεκτικότητα σε σερίνη, προλίνη και γλυκίνη, κάτι που έμεινε χωρίς κάποια *αυστηρή* εξήγηση μέχρι το τέλος, ενώ η αυξημένη περιεκτικότητα σε λευκίνη ίσως δικαιολογείται από την αυξημένη απαίτηση για αυτόν τον τύπο στον υδρόφοβο πυρήνα.

Οι προτιμήσεις, που θεωρητικά αναμένονται, υπάρχουν -τουλάχιστον σε μια πρώτη διαισθητική προσέγγιση. Η συζήτηση επιστρέφει στην τελευταία νύξη, στα πλαίσια της εκτίμησης της σημαντικότητας των διαφορών των σχετικών συστάσεων των επτά θέσεων (σε αμινοξικούς τύπους) από εκείνη των α-ελίκων (ή των πρωτεϊνών γενικά) συνολικά, καθώς και της δυνατότητας των κατανομών των Πινάκων 3 και 4 να διακρίνουν αλληλουχίες που μπορούν να διπλωθούν σαν δεμάτι. Αυτά περιγράφονται σε επόμενο κεφάλαιο, αφού πρώτα διερευνηθούν κάποιες λεπτομέρειες της κατασκευής του δεματιού. Πάντως, καθώς δεν υπάρχει κοινή προέλευση (εξελικτικά) ή κοινή λειτουργία των επτά οικογενειών από τις οποίες προκύπτουν οι πίνακες, οι προτιμήσεις αυτές οφείλουν να αποτελούν μια απεικόνιση σε επίπεδο τύπων αμινοξικών καταλοίπων των περιορισμών που τίθενται λόγω δομής και μόνο.

Οι λέξεις δεν μπορούν να τονίσουν επαρκώς ότι τα αποτελέσματα αυτά είναι στατιστικής φύσεως και αναφέρονται σε προτιμήσεις και όχι απόλυτες επιλογές. Για παράδειγμα, οι θέσεις a και d του υδρόφοβου πυρήνα καταλαμβάνονται από αλανίνη και λευκίνη σε μεγαλύτερο ποσοστό από ότι από άλλους αμινοξικούς τύπους. Αυτό σημαίνει ότι οι περιορισμοί που τίθενται λόγω στερεομετρίας του περιβάλλοντος χώρου, σε συνδυασμό με την απουσία διαλύτη, οδηγούν σε συμβατότητα με αυτούς τους τύπους *σε μεγαλύτερο ποσοστό των περιπτώσεων*. Όμως ***η συμβατότητα μιας συγκεκριμένης θέσης a, με κάποιο συγκεκριμένο αμινοξικό τύπο, δεν μπορεί να καθοριστεί από πριν με ακρίβεια,*** αφού για αυτή τη θέση μπορεί να υπάρχουν συγκεκριμένοι περιορισμοί που δεν γενικεύονται για τις θέσεις a, όπως επαφή με ένα ενεργό κέντρο, ή να αλλάζει σε αντιστάθμιση αλλαγών σε θέσεις γειτονικές στο χώρο. Για παράδειγμα, στις μεσαίες φέτες της πρωτεΐνης ROP υπάρχει μια (και λόγω συμμετρίας του μορίου δύο) ισολευκίνη (Ile15) σε θέση d (όπου συχνότερα απαντά λευκίνη). Ο κενός χώρος που πιθανόν να έμενε, λόγω διαφοράς στη διακλάδωση στην πλευρική αλυσίδα των δύο τύπων, καλύπτεται από τη γειτονική στο χώρο θρεονίνη Thr19 σε θέση a (συχνότερα: αλανίνη ή λευκίνη). Μάλιστα, η Thr19 προσφέρεται, όχι από την αλυσίδα στην οποία βρίσκεται η Ile15, αλλά από τη συμμετρική. Άλλωστε, σε επόμενο κεφάλαιο αναφέρεται η περίπτωση μιας ολόκληρης οικογένειας που έχει κάπως αυξημένο ποσοστό λευκίνης συνολικά για τις θέσεις a, λόγω αυξημένων αποστάσεων μεταξύ των α-ελίκων -και συζητιέται περαιτέρω εκεί. ***Πρόκειται για την "πράξη" της διαλεκτικής αντίθεσης γενίκευση-εξειδίκευση.***

Πάντα στα πλαίσια της συζήτησης για τη στατιστική φύση των αποτελεσμάτων, ίσως πρέπει να διευκρινιστεί τι είδους διαφορά επιφέρει η χρήση δομικής πληροφορίας που συνάγεται από συστοιχίσεις αλληλουχιών. Αυτό φαίνεται καλύτερα αν ξαναγραφτούν οι τύποι των κανονικοποίησεων ως εξής:

$$\mathbf{O}_{f(i,k)} = \{ \sum_{m=1\text{-}Nm} [\mathbf{R}_{m}(i,k) / \sum_{i=1\text{-}20} \mathbf{R}_{m}(i,k)] \} / N_m \text{ ...και...}$$

$$\mathbf{P}_{f(i,k)} = \{ \sum_{m=1\text{-}Nm} [\mathbf{R}_{m}(i,k) / \sum_{k=1\text{-}7} \mathbf{R}_{m}(i,k)] \} / N_m$$

όπου $\mathbf{R}_{m}(i,k)$ είναι ο αριθμός των περιπτώσεων ενός αμινοξικού τύπου ***i*** σε μια θέση ***k***=(a,b..g) για το μέλος ***m*** της οικογένειας ***f***, και N_m ο αριθμός των μελών Εδώ, η κανονικοποίηση γίνεται πρώτα

για κάθε μέλος (!) και μετά τα $\mathbf{O}_f(i,k)$ και $\mathbf{P}_f(i,k)$ (που αφορούν την οικογένεια ολόκληρη) υπολογίζονται σαν μέσοι όροι από τα N_m μέλη. Αυτό, σε τεχνικό επίπεδο και μέσα στα όρια της κάθε οικογένειας, δεν χρειάζεται· γι'αυτό χρησιμοποιήθηκαν (στα όρια της κάθε οικογένειας) οι απλοποιήσεις των τύπων αυτών. Όμως, κάτω από αυτή τη διατύπωση φαίνεται η εξάρτηση της ποιότητας (αξιοπιστίας) της πληροφορίας (που φτάνει στην εξαγωγή των μέσων όρων για όλες τις οικογένειες) από τον αριθμό των μελών, αφού τώρα φαίνεται ότι ο λόγος $\boldsymbol{\sigma}(X)/X \propto 1/\sqrt{N_m}$, όπου X η εκάστοτε προτίμηση ή σχετική σύσταση όπως υπολογίζεται μέσα στην κάθε οικογένεια.

> Η πρακτική εξήγηση είναι η εξής: αν υποτεθεί ότι χρησιμοποιείται μόνο ένα μέλος από την κάθε οικογένεια, τότε η αξιοπιστία της πληροφορίας πχ. για την κάθε θέση d εξαρτάται από το βαθμό συντήρησής της μέσα στην οικογένεια: αν μια θέση d είναι καλά συντηρημένη, τότε οποιοδήποτε μέλος της οικογένειας θα περάσει (με μεγάλη πιθανότητα) την ίδια πληροφορία στην τελική εξαγωγή μέσων όρων για όλες τις οικογένειες, ενώ η πληροφορία, που θα αντιπροσωπεύσει μια λιγότερο καλά συντηρημένη θέση d, θα εξαρτηθεί από το ποιο μέλος της οικογένειας θα επιλεγεί. Αντίθετα, όσο πιο πολλά μέλη χρησιμοποιηθούν (από την κάθε οικογένεια), τόσο η πληροφορία γίνεται ανεξάρτητη από την εκάστοτε συλλογή των μελών. Ακόμα πιο απλά, υποβοηθείται η ανάδειξη χαρακτηριστικών που ισχύουν για ολόκληρες οικογένειες, (ενώ ταυτόχρονα καταστέλλονται εκείνα τα χαρακτηριστικά που αποτελούν ιδιαιτερότητα του κάθε μέλους) με αποτέλεσμα μια εξομάλυνση των στατιστικών λαθών -ζωτικής σημασίας χαρακτηριστικό εκεί που οι αριθμοί των οικογενειών είναι μικροί!
>
> Το κέρδος που μπορεί να έχει κανείς από την επιπλέον πληροφορία έχει δειχτεί και αλλού[18], χωρίς να έχει γίνει επιπλέον συζήτηση. Αυτό το βασικό σημείο της τεχνικής, ήταν κάτι που δεν "συνηθίζοταν" εκείνο τον καιρό, (1990) εκτός από λίγες περιπτώσεις[19]. Η κανονικοποίηση δε μέσα στις οικογένειες πριν εξαχθούν μέσοι όροι είναι κάτι που σπάνια γίνεται ακόμη και σήμερα (1998)· αντίθετα, συνήθως δύο, τρεις ή και περισσότερες αλληλουχίες που ανήκουν σε μια οικογένεια, λαμβάνονται υπ' όψη το ίδιο με αλληλουχίες που είναι μοναδικοί (γνωστοί ή γενικότερα) αντιπρόσωποι των οικογενειών όπου ανήκουν, μετατοπίζοντας τα αποτελέσματα προς την πλευρά των πρώτων.

Ενώ όμως η πληροφορία, που προέρχεται από συστοιχίσεις, απλά αυξάνει την αξιοπιστία της κάθε ανεξάρτητης παρατήρησης, οι επαναλήψεις με την μορφή των επτάδων αυξάνουν δραστικά τον αριθμό των ανεξάρτητων παρατηρήσεων: ακόμη κι αν κάθε οικογένεια είχε μόνο ένα μέλος, για κάθε θέση θα υπήρχαν περισσότερες από 80 παρατηρήσεις. Έτσι, η φράση "στο 31% της θέσης *d* απαντά λευκίνη", θα σήμαινε 25 κατάλοιπα λευκίνης σε θέση *d*, και υποθέτοντας ότι ο αριθμός των παρατηρήσεων κάθε αμινοξικού τύπου στην κάθε θέση ακολουθεί μια κατανομή που προσεγγίζεται σε κάποια όρια από την Poisson, η *αναμενόμενη* τυπική απόκλιση είναι μόλις της τάξης του 5.

[18] Δες για παράδειγμα Rost και Sander (1994) "Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure" *Proteins: Structure Function and Genetics* **19**, 55-72, όπου μόνο με χρήση τέτοιου είδους πληροφορίας σημειώνεται υψηλό ποσοστό επιτυχίας στη μέθοδο πρόβλεψης που περιγράφουν.

[19] Δες Sali A., Overington J.P., Johnson M.S. και Blundell T.L. (1990) "From Comparisons of Protein Sequences and Structures to Protein Modelling and Design" *Trends in Biochemical Sciences*, **15**, 235-240 για ένα από τα σπάνια παραδείγματα.

Κατά τα λοιπά, η πρώτη αυτή ανάλυση ίσως θυμίζει μεθόδους όπως των Chou και Fasman [1978]. Στην πραγματικότητα όμως, όπως και η αντίστοιχη ανάλυση των Lupas et al [1991] για τις υπερ-ελικωμένες έλικες (που όμως δημοσιεύτηκε αφού είχε τελειώσει η παρούσα), έχει μια σημαντική διαφορά: η μέθοδος των Chou και Fasman [1978], αφού υπολογίσει τις προτιμήσεις των 20 αμινοξικών τύπων για α-έλικες και β-πτυχωτά φύλλα, προβλέπει ότι κάθε αμινοξύ μιας αλληλουχίας θα βρεθεί στην δευτεροταγή δομή που προτιμά το ίδιο μαζί με λίγα αμινοξέα γειτονικά στην αλληλουχία (στην πραγματικότητα "ταξινομεί" ένα μικρό τμήμα αλληλουχίας κάθε φορά), χωρίς να λαμβάνει υπ'όψη τη σειρά που ενδεχόμενα θα πρέπει να έχουν αυτά. Έτσι, όπως συζητιέται σε επόμενο κεφάλαιο (κατά τη δημιουργία ενός μοντέλλου βασισμένο στην εργασία που περιγράφεται εδώ), το συνδετικό τμήμα ανάμεσα στις δύο α-έλικες της πρωτεΐνης ROP, καθώς αποτελείται από αμινοξικούς τύπους που απαντούν συχνά σε έλικες, προβλέπεται ελικοειδές, σαν τμήμα μιας συνεχόμενης α-έλικας που καλύπτει σχεδόν όλη την πρωτεΐνη. Όμως, καθώς βγάζει εκτός φάσης την εναλλαγή υδρόφοβων-υδρόφιλων αμινοξέων των ελίκων που συνδέει, αν γινόταν μια τέτοια α-έλικα, μεγάλες υδρόφοβες επιφάνειες θα ήταν σε "διαμετρικά αντίθετες πλευρές του κυλίνδρου", αντί να είναι στην ίδια πλευρά και ευθυγραμμισμένες. Αντίθετα, ***ένα σχήμα πρόβλεψης βασισμένο σε πίνακες όπως οι παραπάνω*** (και όπως και των Lupas et al [1991])**,** δεν αναζητά διαδοχικά αμινοξέα που απλά να είναι συμβατά με το δεμάτι (κάτι που έτσι κι αλλιώς δεν διαφέρει πολύ από το να είναι συμβατά με τις α-έλικες γενικά -Πίνακας 5), αλλά ***αναζητά μια διαδοχή αμινοξέων συμβατή με μια διαδοχή θέσεων***. ***Έτσι, γίνεται εκμετάλλευση μιας εσωτερικής ανακατανομής (στα πλαίσια του προτύπου),*** που εκφράζεται εντονότερα στο υδρόφοβο εσωτερικό του, σαν συνέπεια των πιο συγκεκριμένων απαιτήσεων για τις εκεί θέσεις, και που αργότερα συζητιέται κατά πόσον αποτελεί ίδιον του δεματιού, ή αν παρόμοια δομικά πρότυπα μπορεί να κάνουν παρόμοια εσωτερική ανακατανομή της γενικότερης σύστασης των α-ελίκων σε αμινοξικούς τύπους.

Προσάρτημα: Τα (επιπλέον) προγράμματα που χρησιμοποιήθηκαν

Για να είναι πιο αξιόπιστη η καταμέτρηση, και προπαντός επαναλήψιμη και επεκτάσιμη σε άλλα δομικά πρότυπα, γράφτηκαν κάποια απλά προγράμματα. Ακολουθούν οι λίστες των προγραμμάτων αυτών· το πρώτο, διαβάζει ένα αρχείο σαν αυτό που παρατίθεται για δείγμα και μετρά τον αριθμό των περιπτώσεων κάθε αμινοξικού τύπου σε κάθε θέση, ενώ το δεύτερο διαβάζει τα αποτελέσματα του πρώτου και τα μετατρέπει σε προτιμήσεις. Ένα σχεδόν ταυτόσημο πρόγραμμα, με τις απαραίτητες αλλαγές στο κομμάτι "Μέρος II" κάνει την μετατροπή σε σχετικές συστάσεις. Η τότε μικρή προγραμματιστική εμπειρία του γράφοντος φαίνεται καθαρά· το κέρδος είναι ότι θα πρέπει να είναι προφανές πως δουλεύουν τα προγράμματα αυτά, ακόμη και για λιγότερο πεπειραμένους προγραμματιστές. Όχι λιγότερο όμως, το πρόγραμμα καταμέτρησης ανιχνεύει από μόνο του πόσα διαφορετικά είδη θέσεων υπάρχουν (όπως είναι υποστηρίζει μέχρι 20), ενώ δεν μπερδεύεται όταν υπάρχουν διαφορετικές πρωτεΐνες μέσα στο ίδιο αρχείο. Τέλος, στον πίνακα που γράφεται στο αρχείο που τελικά προκύπτει, τα κενά μεταξύ των αριθμών αντικαθίστανται με στηλοθέτες (tab stops), ώστε να είναι αναγνώσιμο από οποιοδήποτε εμπορικό πρόγραμμα επεξεργασίας πινάκων.

Πρόγραμμα cntpos - ένα απλό πρόγραμμα καταμέτρησης αμινοξικών τύπων που βρίσκονται σε τοπολογικά διακριτές θέσεις

```
#include <stdio.h>
#include <math.h>
#include <string.h>

#define MAXROW 23
#define MAXCOL 20 /* Υποστήριξη μέχρι 20 διαφορετικών θέσεων */

#define BUFSIZ 127
static char aastrng[]="ABCDEFGHIKLMNPQRSTVWXYZ";
static int  matout[MAXROW][MAXCOL];

main() {
 int   r, c, npos, pars;
 int   aatype, toppos;
 char  postyp[MAXCOL]; /* Οι τύποι θέσεων που πραγματικά συναντώνται */
 char  infile[80], outfile[80], lnbuf[BUFSIZ], ttlin[BUFSIZ];
 char  seqbuf[BUFSIZ], radbuf[BUFSIZ];
 FILE  *in, *out;
 int   getln(), remspc(), indx(), xindx();

/* Πρόσβαση στο αρχείο δεδομένων */
 infile[0]='\0';
 printf("\n InFile to use - [RET] to quit: ");
 gets(infile); if (infile[0] == '\0' ) exit(0);
 if ( (in=fopen(infile,"r")) == NULL )
  { printf("\n File not found\n"); exit(1); }

/* Ανάγνωση γραμμής τίτλου */
 if (getln(in,ttlin) == 0-1)
  { printf("\n Empty file \n"); exit(1);}

/* Ανάγνωση δεδομένων - παράλληλη καταμέτρηση */
 postyp[0]='\0';
 while (getln(in,lnbuf) != 0-1) {
  if (xindx(lnbuf,"seqres") == 0 || xindx(lnbuf,"SEQRES") == 0) {

   strcpy(seqbuf,lnbuf);
   while (xindx(lnbuf,"radpos") != 0 && xindx(lnbuf,"RADPOS") != 0) {
    if (getln(in,lnbuf) == 0-1)
     {printf("\n Format error - Missing info for last seqres \n"); exit(1);}
    if (xindx(lnbuf,"seqres") == 0 || xindx(lnbuf,"SEQRES") == 0)
     {printf("\n Format error - Unexpected seqres \n %s\n",lnbuf); exit(1);}
   }/*endwhile*/

   strcpy(radbuf,lnbuf);
   pars=7; /* == στήλη 8 στο αρχείο δεδομένων */
   while (seqbuf[pars] != '\0') {
    if ( (aatype=indx(aastrng,seqbuf[pars])) < strlen(aastrng) ) {
     if ( (toppos=indx(postyp,radbuf[pars])) == strlen(postyp) ) {
      if (toppos > MAXCOL)
       {printf("\n Too many different positions \n"); exit(1);}
      else { postyp[toppos]=radbuf[pars]; postyp[toppos+1]='\0'; }
     }/*endif*/
     ++(matout[aatype][toppos]);
    }/*endif*/
   ++pars; }/*endwhile*/

  }/*endif*/
 }/*endwhile*/
 fclose(in);

/* Δημιουργία αρχείου αποτελεσμάτων */
 npos=strlen(postyp); --npos; printf("\n Npos: %d",npos);
 printf("\n OutFile to contain results - [RET] for screen: ");
 gets(outfile); if (outfile[0] == '\0' ) out=stdout; else {
     if ( (out=fopen(outfile,"w")) == NULL )
      {printf("\n Can not create\n"); exit(1);} }

/* Εγγραφή τίτλου - επικεφαλίδας... */
```

```
  fprintf(out,"%s\n",ttlin);
  for (c=0; c<=npos; ++c ) sprintf(&(lnbuf[c*2]),"\t%c\0",postyp[c]);
  fprintf(out,"%s\n",lnbuf);

/* ...και 23 σειρών αποτελεσμάτων */
  for (r=0; r<=22; ++r) { lnbuf[0] = aastrng[r];
   for (c=0; c<=npos; ++c )
    sprintf( &(lnbuf[c*6+1]),"%6d\0",(int)(*(*(matout+r)+c)) );
   remspc(lnbuf); fprintf(out,"%s\n",lnbuf);
  }/*endfor*/  fclose(out);

 printf("\n"); }/* Endmain */

/* GetLn ***/
int getln(s,b)
FILE *s; char *b; {char *c; c=b;
 while((*c=getc(s))!='\n'){if(*c==EOF){*c='\0';return 0-1;}++c;}*c='\0';
 return 0;}/* End of getln() function */

/* RemSpc ***/
int remspc(b)
char *b; {char *c, *p; c=b;
 while (*c != '\0') {
  if (*c==' ') {
   if ( *(c+1) != ' ') *c ='\t';
   else { p=c; while (*p != '\0') { *p = (*(p+1)); ++p; } continue;}
  }/*endif*/
 ++c;}/*endwhile*/
 return 0;}/* End of remspc() function */

/* Indx ***/
int indx(ptstr,qry)
char *ptstr, qry; { int i=0; char *s;   s=ptstr;
while (*s!='\0'){if(*s==qry) return i; ++i; ++s;} return i;}
/* End of indx() function */

/* XIndx() ***/
int xindx(ptstr,ptqry)
char *ptstr, *ptqry;
{ int i=0, j; char *s,*q;          s=ptstr; q=ptqry;
while (*(s+i) != '\0')
{j=0;while(*(q+j)==(*(s+i+j))){++j;if(*(q+j)=='\0') return i;} ++i;} return -1;}
/* End of xindx() function */
```

Δείγμα αρχείου δεδομένων - κυτόχρωμα β_{562}

```
%>
seqcod 256B
seqres ADLEDNMETLNDNLKVIEKADNAAQVKDALTKMRAAALDAQKATPPKLEDKSPDSPEMKD
radpos ..defgabcdefgabcdefg..abcdefgabcdefgabcdefg.............fgabc

%>
seqcod 256B
seqres FRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYHQKYR
radpos defgabcdefgabcdefgabc..abcdefgabcdefgabcdefga.
```

Πρόγραμμα xnorm - μετατρέπει τα αποτελέσματα του cntpos σε προτιμήσεις.

Οι σχετικές συχνότητες προκύπτουν τροποποιώντας τα αθροίσματα στο Μέρος II.

```
#include <stdio.h>
#include <math.h>
#include <string.h>

#define REAL4 float

#define BUFSIZ 127
static char  aastrng[]="ACDEFGHIKLMNPQRSTVWY";
static char  psstrng[]="abcdefg";
static int   matinp[20][7];
static REAL4 matout[20][7];

main() {
 int matinrow,  matincol;
 int matoutrow, matoutcol;
 int sum, lin, pars, pvtab;
 char tmpchr, testchr;
 char infile[80], outfile[80], lnbuf[BUFSIZ], pslin[BUFSIZ], ttlin[BUFSIZ];
 FILE *in, *out;
 int getln(), remspc();

/*** ΜΕΡΟΣ Ι : Ανάγνωση και έλεγχος αρχείου δεδομένων ***/
/* Πρόσβαση αρχείου δεδομένων */
 infile[0]='\0';
 while( (in=fopen(infile,"r")) == NULL) {
  printf("\n InFile to normalise - [RET] to quit: ");
  gets(infile); if (infile[0] == '\0' ) exit(0);
 }/*endwhile*/

/* Εκτύπωση τίτλου */
 getln(in,ttlin); printf("\n Title : %s",ttlin);

/* Έλεγχος σωστής σειράς των θέσεων οριζόντια*/
 getln(in,lnbuf);
  for (pvtab=1, pars=1, matincol = 0; matincol <= 6; pars++) {
   if ( lnbuf[pars] == '\t' || lnbuf[pars] == '\0' ) {
    tmpchr = lnbuf[pars];
    lnbuf[pars] = '\0';
    if (sscanf(&(lnbuf[pvtab]),"%c",&testchr) != 1) {
     lnbuf[pars] = tmpchr;
     printf("\n Format error in line: \n\n%s\n",lnbuf); exit(1); }
     if (testchr != psstrng[matincol])
      { printf("\n Error in order of positions \n"); exit(1);}
    lnbuf[pars] = tmpchr; pvtab=pars+1; ++matincol;
   }/*endif*/
   if ( lnbuf[pars] == '\0')
    { printf("\n Missing column in line: \n\n%s\n",lnbuf); exit(1); }
  }/*endfor*/
 strcpy(pslin,lnbuf);

/* Ανάγνωση πίνακα */
 for (lin=1; lin<=20; lin++) { matinrow = lin-1;
  getln(in,lnbuf);

/* Έλεγχος σωστής σειράς των θέσεων κάθετα */
  if (lnbuf[0] != aastrng[matinrow])
   {printf("\n Error in order of aminoacids \n"); exit(1);}

/* Ανάγνωση ποσοστού από τη σειρά στον εσωτερικό πίνακα του προγράμματος*/
  for (pvtab=2, pars=2, matincol = 0; matincol <= 6; pars++) {
   if ( lnbuf[pars] == '\t' || lnbuf[pars] == '\0' ) {
    tmpchr = lnbuf[pars];
    lnbuf[pars] = '\0';
    if (sscanf(&(lnbuf[pvtab]),"%d",&(matinp[matinrow][matincol])) != 1) {
     lnbuf[pars] = tmpchr;
     printf("\n Format error in line: \n\n%s\n",lnbuf); exit(1); }
    lnbuf[pars] = tmpchr; pvtab=pars; ++matincol;
   }/*endif*/
   if ( lnbuf[pars] == '\0' )
```

```
    { printf("\n Missing column in line: \n\n%s\n",lnbuf); exit(1); }
  }/*endfor pars*/ }/*endfor lin*/ fclose(in);

/*** ΜΕΡΟΣ ΙΙ : Μετατροπή οριζόντια ***/
 for (lin=1; lin<=20; lin++) { matoutrow = lin-1; sum=0;
  for (matoutcol = 0; matoutcol <= 6; matoutcol++)
   sum += matinp[matoutrow][matoutcol];
  for (matoutcol = 0; matoutcol <= 6; matoutcol++) {
   if (sum != 0)
    matout[matoutrow][matoutcol]=
     (REAL4)(100*matinp[matoutrow][matoutcol])/(REAL4)sum;
   else
    matout[matoutrow][matoutcol]=(REAL4)0;
 }/*endfor matoutcol*/ }/*endfor lin*/

/*** ΜΕΡΟΣ ΙΙΙ : Δημιουργία αρχείου αποτελεσμάτων και εγγραφή ***/
 outfile[0]='\0';
 printf("\n OutFile to contain results - [RET] for screen: ");
 gets(outfile);
 if (outfile[0] == '\0' ) out=stdout;
 else {
  if ( (out=fopen(outfile,"w")) == NULL )
   {printf("\n Unable to create output file \n"); exit(1);}
 }/*endif*/

/* Εγγραφή τίτλου - επικεφαλίδας... */
 fprintf(out,"%s\n",ttlin);
 pslin[15]='\0';
 fprintf(out,"%s\n",pslin);

/* ...και αποτελεσμάτων */
 for (lin=1; lin<=20; lin++) { matoutrow = lin-1;
  lnbuf[0]=aastrng[matoutrow];
  for (matoutcol = 0; matoutcol <= 6; matoutcol++)
   sprintf(&(lnbuf[matoutcol*6+1]),"%6.1f\0",matout[matoutrow][matoutcol]);
  remspc(lnbuf); fprintf(out,"%s\n",lnbuf);
 }/*endfor*/
 fclose(out);

}/* Endmain */

/* GetLn ***/
int getln(s,b)
FILE *s; char *b; {char *c; c=b;
 while((*c=getc(s))!='\n'){if(*c==EOF){*c='\0';return 0-1;}++c;}*c='\0';
 return 0;}/* End of getln() function */

/* RemSpc ***/
int remspc(b)
char *b; {char *c, *p; c=b;
 while (*c != '\0') {
  if (*c==' ') {
   if ( *(c+1) != ' ') *c ='\t';
   else { p=c; while (*p != '\0') { *p = (*(p+1)); ++p; } continue;}
  }/*endif*/
 ++c;}/*endwhile*/
 return 0;}/* End of remspc() function */
```

Μέρος Α / Κεφάλαιο ΙΙ:

Κατανομές κατά μήκος των α-ελίκων και των μεταξύ τους συνδετικών τμημάτων

Εισαγωγή

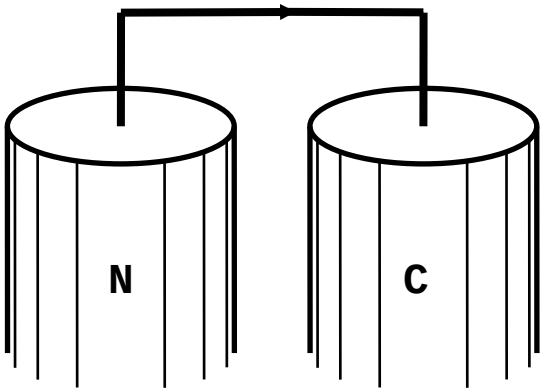
Στο Κεφάλαιο Α.Ι διαπιστώθηκε ένα μοτίβο -αισθητών- διαφορών στις συχνότητες εμφάνισης των διαφόρων αμινοξικών τύπων, σε τοπολογικά διακριτές θέσεις ενός απλού δομικού προτύπου, που μάλιστα ακολουθεί την εσωτερική επανάληψη που ορίζει η τοπολογία του προτύπου αυτού. Το μοντέλλο όμως, που χρησιμοποιήθηκε για την περιγραφή του προτύπου, στην πρώτη αυτή προσέγγιση, είναι απλουστευμένο. Για παράδειγμα, δεν έγινε διάκριση στις περιπτώσεις θέσεων σε ενεργά κέντρα, ούτε λήφθηκε υπ'όψη ότι θέσεις, σε τελικές στροφές α-ελίκων, είναι από τη μία πλευρά εκτεθειμένες στον διαλύτη και έχουν λιγότερους περιορισμούς στο πακετάρισμα, ιδιαιτερότητες που εξηγούν την ύπαρξη φορτισμένων καταλοίπων σε θέσεις a και d, ενώ δείχτηκε πως ογκώδη κατάλοιπα (πχ. τρυπτοφάνης) συχνά χωρούν σε θέσεις a και d κατ' εξαίρεση, μόνο σε ακραίες στροφές. Ακόμη, αγνοήθηκε αυτό που ήταν ήδη γνωστό, ότι οι θέσεις στην πρώτη (κυρίως, αλλά και στην τελευταία) στροφή στις α-έλικες παρουσιάζουν έντονες προτιμήσεις για συγκεκριμένους αμινοξικούς τύπους [Richardson και Richardson, 1988]. Επιπλέον, πρέπει να ληφθεί υπ'όψη ότι τα περισσότερα συνδετικά τμήματα μεταξύ των α-ελίκων, στο δείγμα μας, είναι μικρού μήκους και το ρεπερτόριο των διαμορφώσεων, που μπορούν να έχουν, περιορισμένο [Efimov, 1991] και είναι πιθανό να επηρεάζουν κατά έναν ακόμη τρόπο τις θέσεις αυτές, αφού θα πρέπει -ίσως- να μπορούν να ικανοποιούν και τις συγκεκριμένες διαμορφώσεις.

Με στόχο ένα λεπτομερέστερο μοντέλλο, η παρατήρηση ότι υπήρχαν αρκετά συνδετικά τμήματα με -μόλις- 2 ή 3 κατάλοιπα, ανάμεσα στα τελευταία σε ελικοειδή διαμόρφωση, έδωσε την ιδέα της αναζήτησης κοινών συμπεριφορών σε επίπεδο διαμόρφωσης και -ενδεχόμενα- στατιστικής των θέσεών τους (φυσικά το δεύτερο προϋποθέτει το πρώτο). Επίσης, σε συνδυασμό με τις παραπάνω παρατηρήσεις, έγινε μια προσπάθεια να διαπιστωθούν πιθανές επιδράσεις του μικρού τους μήκους στη συμπεριφορά θέσεων του δεματιού, που -τυχόν- ανήκουν σε ακραίες στροφές. Η εργασία αυτή πραγματοποιήθηκε μέχρι τον Αύγουστο 1992, αλλά η δημοσίευσή της δεν κρίθηκε σκόπιμη (από τον επιβλέποντα), καθώς συμφωνούσε με τις ήδη τότε δημοσιευμένες εργασίες άλλων (που αναφέρθηκαν παραπάνω) και καθώς δεν περιείχε στοιχεία που να ανατρέπουν τα αποτελέσματα του προηγουμένου κεφαλαίου.

Διαδικασία

Σε μια πρώτη φάση εξετάζονται τα συνδετικά τμήματα μεταξύ των α-ελίκων όσον αφορά στερεοδιάταξη και τυχόν προτιμήσεις για έναρξη ή λήξη σε συγκεκριμένες θέσεις a-g. Έπειτα, και με την εικόνα αυτή δεδομένη, δίνονται οι κατανομές των είκοσι αμινοξικών τύπων στις θέσεις των ακραίων στροφών, και κάποιες μεταξύ τους σχέσεις.

Χρήσιμοι ορισμοί: Ένα συνδετικό τμήμα μεταξύ δύο α-ελίκων μπορεί να χαρακτηριστεί σαν δεξιόστροφο ή αριστερόστροφο. Εδώ υιοθετείται ο ορισμός του Efimov [1991], που έχει ως εξής: ας υποθέσουμε ότι κοιτάμε μια α-φουρκέττα (δηλαδή δύο αντιπαράλληλα πακεταρισμένες α-έλικες και το μεταξύ τους συνδετικό τμήμα) από την πλευρά του υδρόφοβου πυρήνα, και έτσι ώστε το συνδετικό τμήμα να βρίσκεται στην κορυφή της σελίδας (Εικόνα 1). Αν η δεύτερη (δηλαδή η καρβοξυτελική) έλικα είναι στα δεξιά, τότε πρόκειται για δεξιόστροφο συνδετικό τμήμα, και αντίστροφα. Επιπλέον, δεν πρέπει να δημιουργείται σύγχυση μεταξύ αυτού του ορισμού της *έννοιας στροφής* και των ορισμών της δεξιό-(αριστερό-)στροφης συστροφής ενός δεματιού και της δεξιό-(αριστερό-)στροφης συνδεσμολογίας του, που δόθηκαν στο Κεφάλαιο Α.Ι. Εξάλλου, σύμφωνα με τους ίδιους ορισμούς [Efimov,1991], η πολυ-πεπτιδική αλυσίδα μπορεί να εισέρχεται στην και να εξέρχεται από την α-έλικα *κάθετα* ή *παράλληλα* σε σχέση με τον κύριο αξονά της τελευταίας.



***Εικ. 1**. Σχηματική επεξήγηση του ορισμού της έννοιας στροφής ενός συνδετικού τμήματος. Αν ο υδρόφοβος πυρήνας βρίσκεται προς την πλευρά του αναγνώστη, τότε σύμφωνα με τον ορισμό που δίνεται στο κείμενο, αυτό το συνδετικό τμήμα είναι δεξιόστροφο.*

Σε όλες τις αναλύσεις που περιγράφονται παρακάτω, εστιάζουμε στο αμινο-τελικό άκρο (για συντομία *N-άκρο*), στο καρβοξυτελικό άκρο (για συντομία *C-άκρο*) και τέσσερις διαδοχικές θέσεις προς το εσωτερικό της α-έλικας από κάθε άκρο (για συντομία *N+1...4* και *C-1...4*). Τα N-άκρα και C-άκρα ορίσθηκαν όπως και στην ανάλυση των Richardson και Richardson [1988] σαν το πρώτο/τελευταίο αμινοξικό κατάλοιπο ο α-άνθρακας του οποίου βρίσκεται "κοντά" (στην παρούσα: όπως διαπιστώνεται οπτικά) στον κύλινδρο της έλικας.

Από τη θέση του N-άκρου και του C-άκρου μπορεί να σχηματίζει (και συνήθως σχηματίζει) έναν υδρογονοδεσμό τύπου α-έλικας, και μπορεί να έχει μια ή και τις δύο δίεδρες γωνίες (φ,ψ) σε διαμόρφωση α-έλικας, αλλά κανένα από τα δύο δεν αποτελεί απαίτηση.

Επιπλέον, κάποιες από τις θέσεις που αναλύονται μπορούν να αποτελούν μέρος του δεματιού, αλλά ούτε αυτό είναι απαραίτητο· γι'αυτό και παρατηρούνται διαφορές στα υπό ανάλυση τμήματα των ελίκων που δείχνονται στον Πίνακα 1 σε σχέση με τον Πίνακα Α.Ι.1.

Οι υπόλοιπες θέσεις των α-ελίκων (που στην ανάλυση των Richardson και Richardson [1988] ονομάζονται απλά "μέση" - middle) χωρίστηκαν σε *Ν-ήμισυ* και *C-ήμισυ*, για να ελεγθούν τυχόν διαβαθμίσεις στις κατανομές ανάμεσα στα δύο μισά. Όπου ο αριθμός ήταν περιττός, το κεντρικό αμινοξικό κατάλοιπο κατατάχθηκε αυθαίρετα στο C-ήμισυ. Επιπλέον, αποφασίστηκε να μην διερευνηθούν θέσεις έξω από τα άκρα, αφού συχνά η πρώτη θέση έξω από το C-άκρο μιας α-έλικας είναι το N-άκρο της επόμενης, εξαιτίας των -γενικά- βραχέων συνδετικών τμημάτων μεταξύ των ελίκων.

Τέλος, σε όλες τις αναλύσεις αναφέρεται σαν *ακτινικά ορισμένη* ή απλά *ακτινική* η θέση a-g και σαν *αξονικά ορισμένη* ή απλά *αξονική* η θέση σε σχέση με το N-άκρο και το C-άκρο.

Δείγμα: Χρησιμοποιήθηκαν οι ίδιες επτά οικογένειες όπως και στο Κεφάλαιο Α.Ι (δες Πίνακα Α.Ι.1). Και εδώ εκμεταλλευτήκαμε όπου ήταν χρήσιμο την υπόθεση ότι η αντιστοίχιση των θέσεων μπορεί να μεταφερθεί με σχετική αξιοπιστία από την αλληλουχία εκείνη που η δομή της είναι πειραματικά προσδιορισμένη σε ομόλογες.

Κάποια νέα μέλη προστέθηκαν, καθώς η αναζήτηση των ομολόγων έγινε σε νεώτερες εκδόσεις των βάσεων δεδομένων NBRF (έκδοση 32.0) και SwissProt (έκδοση 20). Και αντίθετα, αφαιρέθηκαν κάποια μέλη που είχαν προβληματίσει σε τεχνικό επίπεδο, κυρίως επειδή δεν ήταν απόλυτα αξιόπιστη η συστοίχιση με τον αντιπρόσωπο γνωστής δομής. Το νέο δείγμα στην τελική του σύνθεση δίνεται στον Πίνακα 1.

Πίνακας 1.α. *Τα Ν-άκρα και C-άκρα των μορίων γνωστής δομής όπως ορίσθηκαν.*

	Κωδικός PDB	1η α-έλικα	2η α-έλικα	3η α-έλικα	4η α-έλικα
Πρωτεΐνη ROP	1ROP	2T-30D	31A-57G	2T-30D	31A-57G

Φερριτίνη		13H-43R	48L-77G	95S-125N	126N-159G
Αιμερυθρίνη	2HMZ	18Y-39D	40N-66Q	69G-87W	90D-110R
(Μυοαιμερυθρίνη)	(2HMR)	(18Y-39N)	(40S-66K)	(69E-87L)	(92D-115K)
Κυτόχρωμα c’	2CCY	4K-32G	39D-54A	78K-103G	104P-126K
Κυτόχρωμα b_{562}	256B	2D-20A	22N-43A	55S-82G	83K-106R
Λυσοζύμη T4	3LZM	92D-107G	114F-124K	125R-135K	142T-156G
TMV coat protein	2TMV	19D-32G	37T-52W	73N-89T	111T-135G

Πίνακας 1.β. *Ομόλογες αλληλουχίες που χρησιμοποιήθηκαν (σύγκρινε με Πίνακα Α.Ι.2). Με αστερίσκο σημειώνονται οι αντιπρόσωποι κάθε οικογένειας με γνωστή δομή.*

Οικογένεια	Είδος	Κωδικός SwissProt	Κωδ. NBRF	% Ταυτ.
Φερριτίνη				
	Homo sapiens (heavy chain)	FRIHHUMAN	FRHUH	[*]
	Mus musculus (heavy chain)	FRIHMOUSE	S06070	92
	Rattus norvegicus (h. chain)		A39884	94
	Gallus gallus (heavy chain)	FRIHCHICK	A26886	91
	Xenopus laevis (heavy chain)	FRIHXENLA	FRXL	69
	Rana catesbeina (heavy chain)	FRI1RANCA	FRFGL	68
	Rana catesbeina (middle chain)	FRI2RANCA	C27805	66
	Rana catesbeina (light chain)	FRI3RANCA	B27805	61
	Rattus norvegicus (light chain)	FRILRAT	FRRTL	52
	Mus musculus (light chain)		B33355	51
	Homo sapiens (light chain)	FRILHUMAN	FRHUL	56
	Equus caballus (light chain)	FRILHORSE	FRHOL	53
	Oryctolagus cuniculus (l. chain)	FRILRABIT	S01239	55
Αιμερυθρίνη				
[2HMZ]	Themiste dyscritum	HEMTTHEDY	HRTHBD	[*]
[1HRB]	Phascolopsis gouldii	HEMTPHAGO	HRGG	78
	Themiste zostericola	HEMTTHEZO	HRTH	76
	Siphonosoma cumanense	HEMTSIPCU	JT0556	47
Μυοαιμερυθρίνη				
[2MHR]	Themiste zostericola	HEMMTHEZO	HRTHM	[*]
	Nereis diversicolor	HEMMNERDI	S16190	60
Αιμερυθρίνη				[vs.2MHR]
(L. unguis)	Lingula unguis (chain a)	HEMTLINUN	JX0184	42
	Lingula unguis (chain b)	HEMULINUN	JT0560	38
Κυτόχρωμα c’				
[2CCY]	Rhodospirillum molischianum	CYCPRHOMO	CCQFCM	[*]
	Alcaligenes sp. (NCIB 11015)	CYCPALCSP	CCALC	32
	Rhodocyclus gelatinosus	CYCPRHOGE	CCRFCG	32
	Paracoccus sp. (ATCC 12084)	CYCPPARSP	CCPCC8	33
	Rhodobacter sphaeroides	CYCPRHOSH	CCRFCS	26

(c554)	Rhodobacter sphaeroides		S04343	27
	Rhodospirillum rubrum	CYCPRHORU	CCQFCR	29
	Rhodospirillum photometricum	CYCPRHOPH	CCQFCP	30
	Rhodopseudomonas sp.	CYCPRHOSP	CCRFCX	21
	Rhodobacter capsulatus	CYCPRHOCA	CCRFPP	26
	Rhodocyclus tenuis	CYCPRHOTE	CCQFCT	30
	Chromatium vinosum	CYCPCHRVI	CCKRCV	23
TMV coat protein [2TMV]	TMV vulgare	COATTMV	VCTMVU	[*]
	TMV dahlemense	COATTMVDA	VCTMDA	82
	Pepper mild mottle virus	COATPPMVS	JQ1315	71
	TMV U2	COATTMGMV	VCTMU2	70
	TMV ORS	COATTMVOR	VCTMOR	64
	TMV HR	COATTMVHR	VCTMHR	46
	TMV cowpea	COATTMVCO	VCTMCP	42
	Cucumber greenmottle virus	COATCGMVS	JQ1160	36

Εξέταση των συνδετικών τμημάτων: Αρχικά, για κάθε συνδετικό τμήμα, προσδιορίστηκε η έννοια στροφής (δεξιόστροφα ή αριστερόστροφα), σύμφωνα με την Εικόνα 1, με άμεση παρατήρηση σε μια οθόνη γραφικών, και μετρήθηκαν οι δίεδρες γωνίες (φ,ψ) των αμινοξικών καταλοίπων που το απαρτίζουν, χρησιμοποιώντας το πρόγραμμα DSSP [Kabsch και Sander, 1983]. Με βάση αυτές τις δίεδρες γωνίες, κάθε αμινοξικό κατάλοιπο κατατάχθηκε σε μια κατηγορία α, β, γ, δ, α_L ή ε σύμφωνα με την Εικόνα 2/ Γεν.Εισαγωγή.

Επιπλέον, για κάθε συνδετικό τμήμα, ελέγθηκε αν κάποια από τα κατάλοιπά του εκτελούν και χρέη αξονικών ή/και ακτινικών θέσεων. Οι αλληλυπερθέσεις, όπου αναφέρονται, έγιναν χρησιμοποιώντας το πρόγραμμα "O" [Jones et al, 1991] ώστε τα αποτελέσματά τους να είναι άμεσα παρατηρήσιμα στην οθόνη γραφικών, ενώ τοπικοί άξονες α-ελίκων, όπου χρειάστηκαν, υπολογίστηκαν χρησιμοποιώντας μια απλοποιημένη παραλλαγή του αλγορίθμου των Sklenar et al [1989], που εξηγείται στο τέλος του κεφαλαίου.

Ακτινική κατανομή των ελικικών άκρων: Με δεδομένη πλέον την εικόνα των συνδετικών τμημάτων, και σε συνάρτηση με αυτή, είναι χρήσιμο να δει κανείς, αν κάποιες αξονικές θέσεις κοντά στα άκρα συμπίπτουν συστηματικά με κάποιες ακτινικές θέσεις (a-g). Τα N-ήμισυ και C-ήμισυ δεν λαμβάνουν μέρος σε αυτή την ανάλυση, αφού όλες οι ακτινικές θέσεις δοθέντος αρκετού μήκους έλικας, συγκλίνουν στο 1/7.

Πιο συγκεκριμένα, για κάθε οικογένεια ***f*** μετρήθηκε η απόλυτη συχνότητα εμφάνισης κάθε αξονικής θέσης ***j*** σε κάθε ακτινική θέση ***k***. Ο αριθμός αυτός διαιρέθηκε δια τέσσερα, αφού κάθε δεμάτι έχει τέσσερα N-άκρα, τέσσερις θέσεις N+1, κοκ.

Φυσικά σε αυτή τη φάση παίρνει μέρος μόνο ο αντιπρόσωπος γνωστής δομής, αφού όλα τα ομόλογα μόρια υποτίθεται ότι φέρουν την ίδια πληροφορία. Έτσι δημιουργήθηκαν επτά πίνακες $\mathbf{M}_f(j,k)$, ένας για κάθε οικογένεια. Ο μέσος όρος τους υπολογίστηκε ως:

$$\mathbf{M}(j,k) = \sum_{f=1\text{-}7}\mathbf{M}_f(j,k) \;/\; 7$$

Σύσταση των αξονικών θέσεων σε αμινοξικούς τύπους: Σε ένα επόμενο βήμα, προσδιορίστηκε η σχετική σύσταση κάθε αξονικής θέσης στους είκοσι τύπους αμινοξικών καταλοίπων, με τρόπο αντίστοιχο με εκείνο του Κεφαλαίου Α.Ι.

Για κάθε οικογένεια ***f***, μετρήθηκε η απόλυτη συχνότητα $\mathbf{R}_f(i,m)$ του κάθε τύπου ***i*** στην κάθε αξονική θέση/ζώνη ***m***. Η σχετική σύσταση $\mathbf{O}_f(i,m)$ υπολογίστηκε για κάθε οικογένεια χωριστά, ακολουθώντας την κανονικοποίηση:

$$\mathbf{O}_f(i,m) = \mathbf{R}_f(i,m) \;/\; \sum_{i=1\text{-}20}\mathbf{R}_f(i,m)$$

Η μέση σχετική σύσταση ***O(i,m)*** για τις επτά οικογένειες δίνεται από τον τύπο:

$$\mathbf{O}(i,m) = \sum_{f=1\text{-}7}\mathbf{O}_f(i,m) \,/\, 7$$

Σχετική προτίμηση των αμινοξικών τύπων για αξονικές θέσεις: Ο ορισμός ενός μεγέθους όπως η προτίμηση **P** που χρησιμοποιήθηκε στο Κεφάλαιο Α.Ι δεν είναι άμεσος σε αυτή την περίπτωση. Και αυτό γιατί, σε αντίθεση με τις ακτινικά ορισμένες θέσεις, που αντιπροσωπεύουν πάντα το 1/7 του μήκους της α-έλικας, οι αξονικά ορισμένες θέσεις αντιπροσωπεύουν άλλοτε μικρό και άλλοτε μεγάλο τμήμα, ανάλογα με το μήκος της. Έτσι, αν απλά μεταφερθεί ο τύπος που χρησιμοποιήθηκε εκεί, οι προτιμήσεις, που θα υπολογιστούν με αυτό τον τρόπο, εξαρτώνται σε μεγάλο βαθμό από τον λόγο <συχνότητα της θέσης> / <μήκος α-έλικας>. Μια εναλλακτική λύση, είναι να διαιρέσει κανείς την σύσταση **O***(i,m)* με κάποια συχνότητα αναφοράς, π.χ. την μέση σύσταση σε αμινοξικά κατάλοιπα τύπου ***i*** είτε του δεματιού είτε των πρωτεϊνών συνολικά, οπότε προκύπτει η *σχετική προτίμηση* των αμινοξικών τύπων για κάθε αξονική θέση. Ακολουθήθηκε η πρώτη εναλλακτική (μέση σύσταση του δεματιού), κυρίως επειδή συνυφασμένη με την έννοια της προτίμησης είναι η έννοια της κατανομής ενός *δεδομένου* αριθμού από κάθε αμινοξικό τύπο, (παρά ενός *αναμενόμενου*) στις διάφορες θέσεις ενός δομικού προτύπου. Η σύσταση αυτή υπολογίστηκε ξανά ώστε να αντιστοιχεί στο νέο δείγμα.

Σημείωση: Δεν έγινε καμία προσπάθεια, για να βρεθεί αν οι ακτινικές κατανομές επηρεάζονται από την ύπαρξη ενεργών κέντρων, αφού αρκετά από αυτά ήσαν άγνωστα, ενώ σε ορισμένες περιπτώσεις δεν εντοπίζονται καν στο δεμάτι.

Αποτελέσματα

Κατανομή και χαρακτηριστικά των συνδετικών τμημάτων: Μια απογραφή των συνδετικών τμημάτων, ταξινομημένων χονδρικά σύμφωνα με το μέγεθός τους, δίνεται στον Πίνακα 2. Αυτή η -περισσότερο διαισθητική- κατάταξη, έγινε πάνω στην εξής βάση: ως *εξαιρετικά βραχέα* χαρακτηρίστηκαν τα συνδετικά τμήματα όπου υπάρχει ένα μόνο κατάλοιπο μεταξύ των δύο ελίκων, ανάμεσα στα τελευταία τους (ως προς το συνδετικό τμήμα) κατάλοιπα που είναι ακόμη σε ελικοειδή (α- ή 3_{10}) διαμόρφωση· σαν *βραχέα* εκείνα που έχουν δύο ή τρία τέτοια κατάλοιπα· σαν *μέσου μήκους* εκείνα που έχουν τέσσερα ή περισσότερα τέτοια κατάλοιπα, αλλά *δεν διακόπτουν* την συνέχεια του προτύπου· και σαν *μακρά* εκείνα που *διακόπτουν* την συνέχεια του προτύπου, παρεμβάλλοντας σε επίπεδο αλληλουχίας άλλα τμήματα του μορίου, που δομικά δεν σχετίζονται -τουλάχιστον όχι άμεσα- με το δεμάτι[20].

Ένα επαναλαμβανόμενο πρότυπο βραχέος συνδετικού τμήματος που εντοπίζεται με την πρώτη ματιά από τον Πίνακα 2, είναι εκείνο που απαντά σαν δεξιόστροφο στην πρωτεΐνη ROP, και στη φερριτίνη μεταξύ 3ης και 4ης α-έλικας, και σαν αριστερόστροφο στο κυτόχρωμα b_{562} επίσης μεταξύ 3ης και 4ης α-έλικας, και στη λυσοζύμη του φάγου T4 μεταξύ 2ης και 3ης α-έλικας. Σύμφωνα με τους ορισμούς του Efimov [1991] απαρτίζεται από μια *κάθετη έξοδο* από την αμινοτελική α-έλικα και μια *κάθετη είσοδο* στην καρβοξυτελική.

> Σε μεγαλύτερη λεπτομέρεια, κατά μήκος της αλυσίδας απαντά κανείς με τη σειρά τα εξής (Εικόνα 2): το τελευταίο κατάλοιπο της πρώτης έλικας με το κανονικό (ελικοειδές) πρότυπο υδρογονοδεσμού και σε διαμόρφωση γ· το C-άκρο της σε διαμόρφωση α_L, που μετατοπίζει την αλυσίδα μακριά από την πρώτη α-έλικα και προς τη δεύτερη· ένα κατάλοιπο που εκτελεί την αντιστροφή της κατεύθυνσης της αλυσίδας σε σχέση με τον άξονα της πρώτης α-έλικας, και ταυτόχρονα αποτελεί το N-άκρο της δεύτερης με δίεδρες γωνίες $(\varphi,\psi)=(-90,+90)$· και τέλος το πρώτο κατάλοιπο της δεύτερης α-έλικας σε ελικοειδή διαμόρφωση και αντίστοιχο πρότυπο υδρογονοδεσμού.
>
> Όπως φαίνεται και στην Εικόνα 2, αυτή η κατασκευή προβάλλει τα β-άτομα άνθρακα από δύο αμινοξικά κατάλοιπα προς την πλευρά του υδρόφοβου πυρήνα και από τα άλλα δύο προς την αντίθετη. Στα δεξιόστροφα συνδετικά τμήματα (όπως στις περιπτώσεις της πρωτεΐνης ROP και της φερριτίνης) οι δύο θέσεις που προβάλλουν προς τον υδρόφοβο πυρήνα είναι το τελευταίο κατάλοιπο της

[20] Τα μακρά συνδετικά τμήματα δεν δείχνονται στον Πίνακα 2, και είναι τα εξής: μεταξύ 2ης και 3ης α-έλικας στην φερριτίνη· μεταξύ 2ης και 3ης α-έλικας στα κυτοχρώματα c' και b_{562}· και τέλος τα συνδετικά τμήματα μεταξύ 2ης και 3ης και μεταξύ 3ης και 4ης α-έλικας στην πρωτεΐνη του καλύμματος του ιού της μωσαϊκής του καπνού.

πρώτης α-έλικας και το Ν-άκρο της δεύτερης (το κατάλοιπο όπου γίνεται και η αντιστροφή της κατεύθυνσης της αλυσίδας), δηλαδή η πρώτη και η τρίτη θέση αυτής της κατασκευής.

Πίνακας 2α. *Περιγραφή των βραχέων συνδετικών τμημάτων που απαντώνται στο δείγμα, σε όρους διαμόρφωσης κύριας αλυσίδας, ακτινικών θέσεων και συντηρητικότητας. Δίνονται δεδομένα για το τμήμα από το τελευταίο κατάλοιπο της πρώτης α-έλικας που βρίσκεται ακόμη σε ελικοειδή διαμόρφωση, μέχρι το πρωτο αντίστοιχο της δεύτερης.*

Κατάλοιπο	φ	ψ	Κατάταξη	Ακτ. θέση[21]	Σε ομόλογες θέσεις:
[α. Εξαιρετικά βραχέα]					
Κυτόχρωμα c' (3-4)					
(αριστερόστροφο)					
102A	-80.2	-29.1	γ/α	c	ATDATTATATT
103G	135.5	169.4	E	-	GGGGDGGGGGG
					DDE--DDEDDE
					--E--------
104P	-58.5	-40.1	A	a	LFFQLKTAAAA
[β. Βραχέα]					
[β1. Δεξιόστροφα]					
Rop (1-2)					
29L	-89.8	-0.7	Γ	d	
30D	57.3	37.7	α_L	-	
31A	-94.2	89.1	B	"a"	
32D	-59.2	-54.8	A	b	
Φερριτίνη (3-4)					
124K	-91.0	3.5	Γ	d	KKKKKKKRRRQH
125N	52.5	61.1	α_L	-	NNNVVVSTTTAT
126D	-121.3	87.1	B	"a"	DDDDDDDDDDDD
127P	-61.7	-34.0	A	b	PPPPPPPPPPPP
[β2. Αριστερόστροφα]					
Λυσοζύμη T4 (2-3)					
123Q	-83.0	-4.5	Γ	c	
124K	66.3	33.4	α_L	-	
125R	-97.2	77.6	β/δ	-	
126W	-51.9	-53.4	A	a	
Κυτόχρωμα b_{562} (3-4)					
81E	-93.8	8.5	Γ	c	
82G	78.6	9.8	α_L	-	
83K	-93.0	70.8	Δ	-	
84V	-62.8	-53.0	A	a	
Κυτόχρωμα b_{562} (1-2)					
19K	-97.1	0.7	Γ	f	

[21] Όταν μια ακτινική θέση (a-g) δίνεται σε εισαγωγικά, αυτό σημαίνει ότι το αντίστοιχο αμινοξικό κατάλοιπο δεν είναι σε ελικοειδή διαμόρφωση, αλλά πιστοποιείται ο ρόλος του σε χρέη κάποιας ακτινικής θέσης με βάση άλλα κριτήρια, πχ. την θέση της πλευρικής του αλυσίδας.

20A	-60.6	144.7	B	“g”
21D	-104.5	-18.1	Γ	-
22N	-146.9	177.6	B	-
23A	-69.4	-38.1	A	a

(Συνεχίζεται)

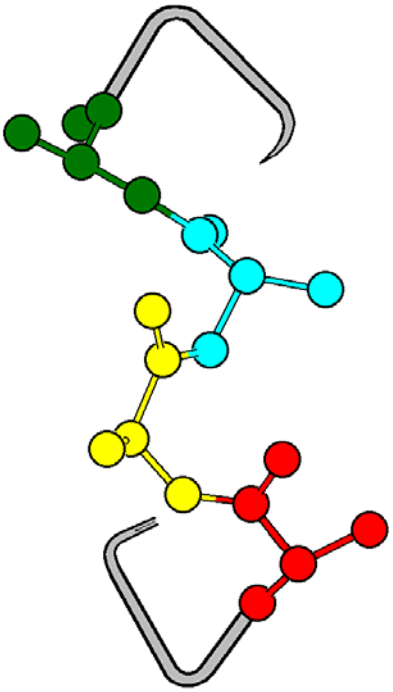
Πίνακας 2α. *Συνέχεια.*

Κατάλοιπο	φ	ψ	Κατάταξη	Ακτ. θέση	Σε ομόλογες θέσεις:
Αιμερυθρίνη (1-2)					
37Q	-79.1	-40.1	A	f	IIN
38A	-160.2	112.1	B	"g"	DDV
39D	-79.8	82.8	β/δ	-	DDG
40N	-127.9	-178.0	B	-	NNG
41A	-69.8	-41.2	A	e	AAA
Μυοαιμερυθρίνη (1-2)[22]					
37R	-69.9	-38.2	A	f	G EE
38D	-150.0	94.9	B	"g"	G FF
39N	-79.6	67.9	Δ	-	N NN
40S	-111.6	165.7	B	-	N TT
41A	-61.5	-45.2	A	e	A RQ

(Τέλος Πίνακα 2α)

Σε αυτή την περίπτωση, η πρώτη θέση είναι ταυτόχρονα (ακτινική) θέση d στο δεμάτι, ενώ η τρίτη εκτελεί χρέη (ακτινικής) θέσης a, χωρίς μάλιστα να έχει διαμόρφωση έλικας. Αντίθετα, στα αριστερόστροφα συνδετικά τμήματα του είδους, που απαντούν στο κυτόχρωμα b_{562} και στη λυσοζύμη του φάγου T4, στον υδρόφοβο πυρήνα προβάλλουν τα β-άτομα άνθρακα της δεύτερης και της τέταρτης θέσης. Το πρώτο και το τέταρτο κατάλοιπο έχουν (ακτινικές) θέσεις c και a, ενώ το N-άκρο της δεύτερης έλικας *δεν θεωρείται ότι είναι* σε θέση g.

[22] Σαν ομόλογες για την μυοαιμερυθρίνη του *Themiste dyscritum* δίνονται η μυοαιμερυθρίνη της *Nereis diversicolor* και οι αλυσίδες α και β της αιμερυθρίνης της Lingula unguis (δες και Πίνακα 1)



***Εικ. 2**. Σχηματική αναπαράσταση του συχνότερα απαντώμενου τύπου συνδετικού τμήματος. Οι διαδοχικές θέσεις δίνονται με διαφορετικά χρώματα. Αν ο υδρόφοβος πυρήνας βρίσκεται προς τα δεξιά, τότε το πρώτο κατάλοιπο (κόκκινο-κάτω μέρος) ακτινικά είναι σε θέση d, ενώ το τρίτο (κυανό) τοποθετείται έτσι ώστε να εκτελεί χρέη θέσης a, αν και δεν έχει καν διαμόρφωση α-έλικας. Αν ο υδρόφοβος πυρήνας βρίσκεται προς τα αριστερά, τότε το πρώτο κατάλοιπο είναι σε θέση c, με το τέταρτο (πράσινο-άνω μέρος) σε θέση a, ενώ το τρίτο δεν ανήκει πλέον στο δεμάτι (δεν θεωρείται δηλαδή ότι είναι σε θέση g).*

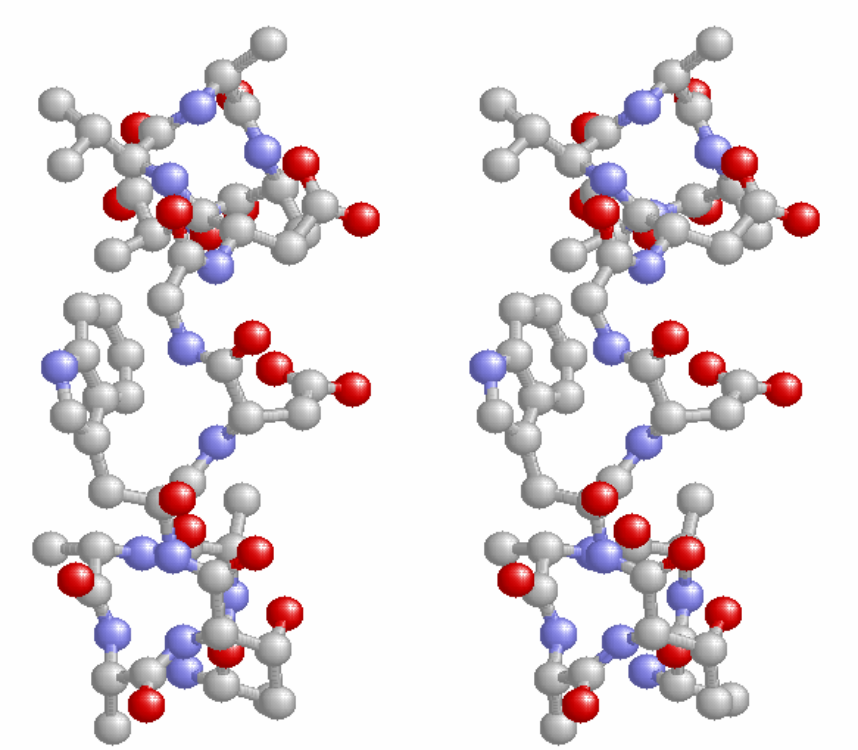
Αυτή η κατασκευή απαντά, όχι μόνο σαν ολοκληρωμένο συνδετικό τμήμα, αλλά και σαν επιμέρους τμήμα σε μεγαλύτερες κατασκευές. Έτσι, το συνδετικό τμήμα μεταξύ 2ης και 3ης α-έλικας στην αιμερυθρίνη και μυοαιμερυθρίνη, και -πιο εμφανώς- εκείνο μεταξύ 3ης και 4ης στην αιμερυθρίνη (Εικόνα 3), εμπεριέχουν αυτή την κατασκευή. Οι αποστάσεις μεταξύ των αξόνων των "γειτονικών" α-ελίκων, στο ύψος των συνδετικών τμημάτων, είναι 10.4, 11 και 12 Ε αντίστοιχα, ενώ π.χ. στην πρωτεΐνη ROP 8.7 Ε.

Το εξαιρετικά βραχύ συνδετικό τμήμα μεταξύ 3ης και 4ης α-έλικας στο κυτόχρωμα c', μπορεί να θεωρηθεί σαν μια επιμέρους περίπτωση του ίδιου τύπου, όπου το ίδιο αμινοξικό κατάλοιπο, η γλυκίνη στη θέση 103, εκτελεί χρέη C-άκρου, N-άκρου, και αντιστροφής της κατεύθυνσης. Αυτό, αν δεν είναι λάθος κατά τον προσδιορισμό της αλληλουχίας ή της δομής, γίνεται εφικτό χάρη στην εκτεταμένη διαμόρφωση αυτής της γλυκίνης, και την πολύ μικρή απόσταση μεταξύ των δύο συγκεκριμένων ελίκων (7.9 Ε), που με τη σειρά της επιτυγχάνεται επειδή ανάμεσα τους πακετάρονται μόνο κατάλοιπα αλανίνης. Ας σημειωθεί, πάντως, πως οι περισσότερες ομόλογες αλληλουχίες φαίνονται (από τις συστοιχίσεις) να έχουν εισδοχή ενός αμινοξικού καταλοίπου στην ίδια θέση.

Αντίθετα, το συνδετικό τμήμα μεταξύ 2ης και 3ης α-έλικας στην αιμερυθρίνη και μυοαιμερυθρίνη, κατατάσσεται στα μέσου μεγέθους, αν και φαινομενικά ο συνδυασμός "κλειδί" α_L-β (66Q-67Y στην αιμερυθρίνη) περιβάλλεται από κατάλοιπα σε ελικοειδή διαμόρφωση· από αυτά όμως, τα 68A-69G δεν ανήκουν στην έλικα που ακολουθεί, αλλά στο συνδετικό τμήμα, το οποίο μεταβάλλεται έτσι ώστε να έχει παράλληλη είσοδο σ'αυτή την έλικα, αντί για την κάθετη που -διαφορετικά- θα αναμένοταν.

Τα μέσου μεγέθους συνδετικά τμήματα δεν παρουσιάζουν κάποια άμεσα ορατή επαναληψιμότητα σε επίπεδο συνολικής κατασκευής, αν και τμήματά τους, όπως έξοδοι από και είσοδοι σε α-έλικες παρουσιάζουν κάποια κοινά μοτίβα. Έχει ήδη σχολιασθεί το γεγονός [Efimov, 1991] ότι καθώς το μέγεθός τους αυξάνει, πολλαπλασιάζονται οι πιθανές διαμορφώσεις και ίσως "απλά" να απαιτείται μεγαλύτερο δείγμα για να παρουσιαστεί κάποια επαναληψιμότητα. Μια περίπτωση αξίζει σχολιασμό: στη λυσοζύμη του φάγου T4, τα συνδετικά τμήματα μεταξύ 1ης και 2ης, και μεταξύ 3ης και 4ης α-έλικας, σχηματίζουν τα ίδια μικρές α-έλικες (Εικόνα 4). Θα μπορούσε κανείς να τα κατατάξει στα μακρά συνδετικά τμήματα λόγω μεγέθους, αλλά συμπεριλήφθηκαν με τα μέσου μεγέθους, επειδή δεν διακόπτουν την συνέχεια του προτύπου.

Μακρά συνδετικά τμήματα, που διακόπτουν την συνέχεια του δεματιού, έχουν συνολικές δομές ειδικές κατά οικογένεια, και στο παρόν δείγμα παίζουν ειδικούς ρόλους: όπως φαίνεται εξετάζοντας τα αντίστοιχα αρχεία από την PDB, στην πρωτεΐνη του καλύμματος του ιού της μωσαϊκής του καπνού συμμετέχει στην συναρμολόγηση του ιού, ενώ στα κυτοχρώματα c' και b_{562} συμμετέχει στον σχηματισμό της τσέπης για την αίμη· και εδώ ένα τμήμα οργανώνεται σε μια μικρή α-έλικα.



Εικ. 3. *Στερεοσκοπική απεικόνιση του (μέσου μήκους) συνδετικού τμήματος μεταξύ των α-ελίκων 3 και 4 της αιμερυθρίνης (κωδ. PDB:2HMZ). Ο προσανατολισμός είναι ίδιος με εκείνο της Εικόνας 2: η αμινοξική έλικα κάτω (φαίνεται το C-άκρο της) και η καρβοξυτελική πάνω (φαίνεται το N-άκρο της). Τα κατάλοιπα του συνδετικού τμήματος μεταξύ των α-ελίκων δείχνονται πλήρη (με τις πλευρικές αλυσίδες), ενώ από εκείνα των α-ελίκων δείχνεται μόνο η κύρια αλυσίδα (μέχρι το β-άτομο άνθρακα). Καθώς το τμήμα είναι αριστερόστροφο, ο υδρόφοβος πυρήνας είναι αριστερά. Η τρυπτοφάνη στο μέσο της αριστερής πλευράς είναι η ίδια που δείχνεται στην Εικόνα Α.Ι.4. Μαζί με την θρεονίνη που προηγείται (κάτω δεξιά -ας προσεχτεί η τοποθέτηση της πλευρικής της αλυσίδας ώστε να καλύπτονται δεσμοί υδρογόνου της κύριας αλυσίδας) αποτελούν τα "επιπλέον" κατάλοιπα (σε σχέση με ένα συνδετικό τμήμα σαν εκείνο της Εικόνας 2). Εαν δεν υπήρχαν, ο πεπτιδικός δεσμός που προηγείται της θρεονίνης θα συνέπιπτε με εκείνον που ακολουθεί την τρυπτοφάνη (ας προσεχτεί ότι έχουν τον ίδιο προσανατολισμό). Είναι λοιπόν προφανής ο τρόπος με τον οποίο αυξάνει η απόσταση μεταξύ των α-ελίκων στα 12 Ε (σε σχέση πχ με τα 8.7 στην πρωτεΐνη ROP). Κατά τα λοιπά το ασπαρτικό οξύ που ακολουθεί την τρυπτοφάνη (και φαίνεται στο μέσο της δεξιάς πλευράς) είναι σε διαμόρφωση γ, η επόμενη γλυκίνη σε διαμόρφωση α_L, (και αντιστοιχεί στη 2η θέση της Εικόνας 2, την "κίτρινη") και το*

ασπαρτικό οξύ (πάνω δεξιά) αποτελεί το Ν-άκρο της καρβοξυτελικής έλικας σε διαμόρφωση β (ας προσεχτεί πάλι η τοποθέτηση της πλευρικής της αλυσίδας).

Πίνακας 2β. *Περιγραφή των μέσου μήκους συνδετικών τμημάτων που απαντώνται στο δείγμα. Δες και Πίνακα 2α για περισσότερες λεπτομέρειες*

Κατάλοιπο	φ	ψ	Κατάταξη	Ακτ.θέση	Σε ομόλογες θέσεις:
[1. Δεξιόστροφα]					
Αιμερυθρίνη (2-3)					
64A	-62.8	-24.9	γ/α	f	AAV
65S	-113.8	14.6	Γ	"g"	SSA
66Q	49.9	43.8	α_L	-	QQK
67Y	-51.3	126.7	B	-	YYY
68A	-63.4	-26.7	γ/α	-	QQG
69G	-94.3	14.4	Γ	"g"	FFG
70Y	-53.8	-52.4	A	a	YYY
Μυοαιμερυθρίνη (2-3)					
64A	-58.8	-40.7	A	f	A KR
65A	-83.1	-0.3	Γ	"g"	S AS
66K	48.4	46.0	α_L	-	A NN
					A --
67Y	-62.0	133.4	B	-	Y YY
					K --
68S	-70.9	-25.3	γ/α	-	S EV
69E	-109.3	29.2	γ/δ	"g"	E HN
70V	-70.2	-28.1	A	a	H FT
Φερριτίνη (1-2)					
42D	-85.6	-8.6	Γ	b	DDDDDDNDDDDD
43R	-61.9	147.0	B	"c"	RRRRRRRRRRR
44D	-65.5	-22.1	Γ	-	DDDDDDDDDDD
45D	-97.7	13.3	Γ	-	DDDDDDDDDDD
46V	-124.7	-63.4	Δ	-	VVVVIVVVVVV
47A	46.9	58.4	α_L	-	AAAAAAAAAAA
48L	-115.5	90.6	B	"a"	LLLLLLLLLLL
49K	-66.2	-27.0	γ/α	b	KKKHHHSEEEE
Λυσοζύμη T4 (1-2)					
106M	-119.5	-1.2	Γ	e	
107G	83.1	149.7	E	-	
108E	-49.9	-51.1	A	-	
109T	-59.1	-46.6	A	-	
110G	-58.8	-49.2	A	-	
111V	-66.3	-41.4	A	-	
112A	-57.1	-23.2	Γ	-	
113G	-71.0	-17.7	Γ	-	
114F	-81.8	54.1	Δ	-	
115T	-43.0	-56.3	A	a	

(Συνεχίζεται)

Πίνακας 2β. *Συνέχεια.*

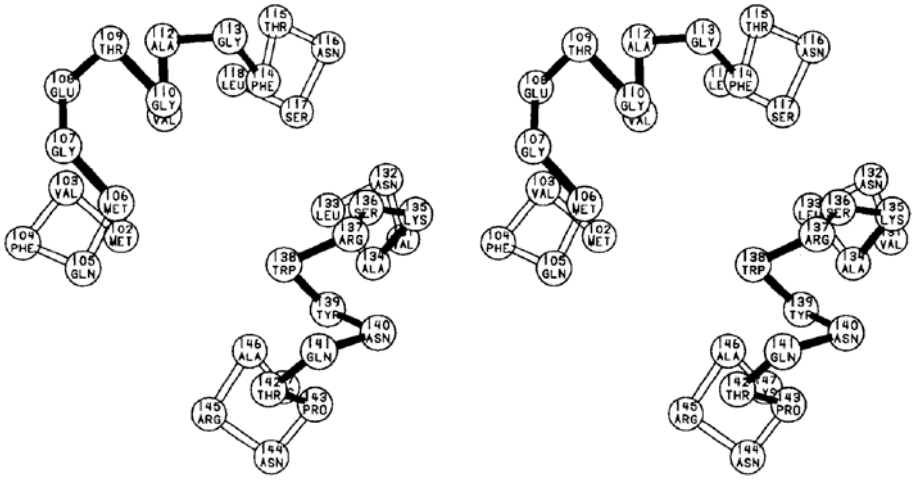
Κατάλοιπο	φ	ψ	Κατάταξη	Ακτ.θέση	Σε ομόλογες θέσεις:
Λυσοζύμη T4 (3-4)					
134A	-72.1	-15.3	Γ	b	
135K	-92.3	67.2	Δ	-	
136S	-151.3	157.8	B	-	
137R	-58.7	-40.9	A	-	
138W	-46.2	-51.1	A	-	
139Y	-65.9	-47.3	A	-	
140N	-64.7	-33.4	A	-	
141Q	-80.6	-40.1	A	-	
142T	-133.1	66.3	Δ	-	
143P	-59.0	-39.5	A	e	
[2. Αριστερόστροφα]					
TMV coat prot. (1-2)					
31L	-54.7	-10.0	Γ	e	LLLLLRQ
32G	-109.1	81.5	β/δ	"f"	GGGGSSG
33N	-165.0	-161.9	B	-	NNNNQNT
34Q	-145.8	116.9	B	-	QQQQSSA
35F	-85.9	21.2	Γ	-	FFFFYFF
36Q	-125.3	10.4	Γ	-	QQQQQQQ
37T	-110.6	104.2	B	-	TTTTTTT
38Q	-44.1	-27.7	γ/α	-	QQQQQQQ
39Q	-69.9	-70.5	A	b	QQQQASA
Αιμερυθρίνη (3-4)					
85D	-74.9	-11.5	Γ	b	DDK
86T	-123.2	33.9	δ/γ	"c"	NNG
87W	-49.5	123.8	B	"d"	WWG
88D	-109.5	-0.7	Γ	-	KKS
89G	84.6	14.6	α_L	-	GGA
90D	-85.2	85.9	Δ	-	DDD
91V	-67.2	-34.4	α/γ	a	VVA
Μυοαιμερυθρίνη (3-4)					
85G	-62.2	-28.6	γ/α	b	R GR
86G	-93.0	-9.7	Γ	c	G HG
87L	-76.7	160.4	B	"d"	L WW
88S	-137.3	147.8	B	-	S KQ
89A	-82.9	142.3	B	-	A AS
90P	-76.6	135.1	B	-	P PP
91V	-67.1	137.8	B	-	V VV
92D	-71.2	156.6	B	-	P PP
93A	-52.9	-35.9	A	e	N QQ

(Συνεχίζεται)

Πίνακας 2β. *Συνέχεια.*

Κατάλοιπο	φ	ψ	Κατάταξη	Ακτ.θέση	Σε ομόλογες θέσεις:
Κυτόχρωμα c' (1-2)					
31A	-87.4	-18.9	Γ	c	KQQKKKKKDE
32G	91.9	10.6	α_L	-	GGGGEEGNSSG
33K	-96.3	-4.1	Γ	-	QKDEEDQ----
34A	-156.8	167.5	B	-	AAIMTLL--P-
35D	-81.3	170.3	B	-	PPEPPAP--QE
36L	-69.2	131.6	B	-	YFYYYVLYFYY
					DDDDD--DDNN
37P	-71.1	134.5	B	-	AAAAAPNAAKA
38A	-68.0	-25.5	γ/α	-	AKDAENAEEDA
39D	-106.0	27.8	δ/γ	c	QVEAVQEAADQ
40A	-55.7	-32.5	A	d	IAFAATAAAGV

(Τέλος Πίνακα 2β)



Εικ. 4*. Στερεοσκοπική σχηματική αναπαράσταση των μέσου μήκους συνδετικών τμημάτων της λυσοζύμης του φάγου T4, που αναφέρονται στο κείμενο ότι σχηματίζουν τα ίδια μικρές α-έλικες.*

<u>*Σχέση μεταξύ αξονικών και ακτινικών θέσεων:*</u> Η κατανομή κάθε αξονικά ορισμένης θέσης, στις επτά διαφορετικές ακτινικά ορισμένες θέσεις, δείχνεται στον Πίνακα 3. Σε 36%, των περιπτώσεων, η α-έλικα ξεκινά με την θέση N+1 σε θέση b, την N+2 σε θέση c, την N+3 σε θέση d, και την N+4 σε θέση e. Στο 74% αυτού του 36% (δηλαδή 25% του συνόλου), το N-άκρο τοποθετεί την πλευρική του αλυσίδα έτσι ώστε να εκτελεί χρέη θέσης a, ενώ στα υπόλοιπα υπάρχει κάποια άλλη διευθέτηση. Επίσης, σε 19% των N-άκρων, οι θέσεις N+1…4 είναι σε ακτινικές θέσεις a-b-c-d. Στον ίδιο πίνακα εμφανίζονται προτιμήσεις για κάποιες ακτινικές θέσεις και στο C-άκρο: στο 28% των περιπτώσεων, η α-έλικα τερματίζει με την θέση

C-4 σε θέση g, την C-3 σε θέση a, την C-2 σε θέση b, και την C-1 σε θέση c, και σε ακόμη 21% με τις θέσεις C-4…1 σε ακτινικές θέσεις a-b-c-d. (Η διαφορετική τιμή της θέσης C-1 οφείλεται σε ανωμαλίες στο πρότυπο της επτάδας στα άκρα μερικών α-ελίκων.)

Πίνακας 3. *Κατανομή (%) των αξονικών θέσεων στις επτά ακτινικές θέσεις a-g.*

	a	b	c	d	e	f	g	Διάφορα
N-άκρο	25	0	4	0	0	0	7	64
N+1	19	36	0	14	13	14	0	4
N+2	0	19	36	0	14	13	14	4
N+3	14	0	19	36	0	14	13	4
N+4	13	14	0	19	36	4	14	0
C-4	21	11	11	7	11	11	28	0
C-3	28	21	11	11	7	11	11	0
C-2	11	28	21	4	11	7	11	7
C-1	11	11	28	25	4	11	7	4
C-άκρο	0	0	7	11	0	0	11	71

Οι συνδυασμοί αξονικών/ακτινικών θέσεων, που αναφέρθηκαν, φέρνουν τα N/C-άκρα μιας έλικας στην εγγύς προς μια γειτονική έλικα πλευρά της, διευκολύνοντας -*τυχόν*- σχηματισμό βραχέος συνδετικού τμήματος. Ελέγχθηκε, λοιπόν, μήπως κάποια N/C-άκρα, με αυτούς τους συνδυασμούς χαρακτηριστικών, προηγούνται ή έπονται συστηματικά βραχέων συνδετικών τμημάτων, και -άρα, ίσως- αυτοί οι συνδυασμοί αποτελούν απόρροια της επίδρασης του μικρού μήκους τους. Από τον Πίνακα 2, προκύπτει ότι τα πρότυπα αυτά απαντούν πριν/μετά και από σύντομα και από μακρύτερα συνδετικά τμήματα. Επιπλέον, η κατανομή τους είναι ομοιόμορφη στις επτά οικογένειες: για παράδειγμα, όλες οι οικογένειες εκτός από το κυτόχρωμα β_{562} περιλαμβάνουν τουλάχιστον μια α-έλικα με τις αμινοτελικές θέσεις N+1...4 σε ακτινικές θέσεις b-c-d-e (στη φερριτίνη τρεις). Αντίθετα, οι χαμηλές εως μηδενικές εμφανίσεις τμημάτων που απολήγουν με τη θέση N+1 σε ακτινική θέση c ή g, και αντίστοιχα τη θέση C-1 σε θέση e, συμπίπτουν με το γεγονός ότι θα είχαν το N/C-άκρο, όχι στην εγγύς προς τη γειτονική έλικα πλευρά τους, αλλά κάπου στην αντίθετη.

Σχετική σύσταση και σχετική προτίμηση κατά μήκος των α-ελίκων: Οι μέσοι όροι από τις επτά οικογένειες, για τα δύο αυτά μεγέθη, δίνονται στους Πίνακες 4 και 5 αντίστοιχα. Η σχετική σύσταση κάθε αξονικής θέσης και η σχετική προτίμηση των αμινοξικών τύπων για κάθε θέση πρέπει να χρησιμοποιούνται μαζί, προκειμένου να εξάγονται ασφαλή συμπερά-

σματα: μια έντονη σχετική προτίμηση ενός αμινοξικού τύπου για μια θέση μπορεί να μην σημαίνει τίποτε, αν η σύσταση της θέσης στον τύπο αυτό είναι μηδαμινή, και αντίστροφα.

> Οι αμινοξικοί τύποι έχουν τοποθετηθεί έτσι ώστε -χονδρικά- να βρίσκονται πιο κοντά στην κορυφή εκείνοι με τους αριθμητικά μεγαλύτερους συνδυασμούς σχετικής σύστασης-σχετικής προτίμησης για το N-άκρο. Για ποσοστά χαμηλότερα του 10-12% στον Πίνακα 4, πρέπει να θεωρείται ότι μπορούν να αλλάξουν κατά πολύ με τη χρήση κάποιου μεγαλύτερου δείγματος.

Αρχικά, παρατηρεί κανείς ότι, στις ακραίες στροφές, για κάθε αξονικά ορισμένη θέση υπάρχουν 2-3 αμινοξικοί τύποι που συγκεντρώνουν περισσότερο από 40%. Αντίθετα, στο N-ήμισυ και C-ήμισυ, παρατηρείται μια ομαλότερη κατανομή. Η αλανίνη και η λευκίνη, που ξεχώρισαν στο Κεφάλαιο A.I, δείχνουν συνολικά μια ομοιόμορφη κατανομή κατά μήκος του δεματιού, με μια τάση -πιο έντονη για τη λευκίνη- να αποφεύγουν κάποιες θέσεις στις ακραίες στροφές και να προτιμούν κάποιες άλλες. Από την άλλη, η προλίνη και η γλυκίνη που αναφέρθηκαν σαν δύο σπάνιοι τύποι στις ακτινικές κατανομές, εδώ διαπιστώνεται ότι έχουν τους "συνήθεις" πολύ συγκεκριμένους ρόλους κατά μήκος των ελίκων. Η προλίνη απαντά σε -περίπου- 20% των θέσεων N+1, δείχνοντας την υψηλότερη σχετική προτίμηση, ενώ η γλυκίνη καταλαμβάνει το 35% των C-άκρων, που αντιστοιχεί στην δεύτερη υψηλότερη σχετική προτίμηση. Άλλες έντονες παρουσίες στον Πίνακα 4 είναι το ασπαρτικό οξύ, η θρεονίνη (και σε κάποιο βαθμό η σερίνη) στο N-άκρο, το γλουταμικό οξύ και η γλουταμίνη στις θέσεις N+2, N+3 και C-2, ενώ τα επίσης υδρόφιλα ασπαραγίνη, λυσίνη και αργινίνη καταλαμβάνουν πιο "ήπια", αλλά πάντως σχετικά υψηλά ποσοστά σε διάφορες θέσεις κατά μήκος του δεματιού.

Πίνακας 4. *Σύσταση (%) των αξονικών θέσεων στους είκοσι αμινοξικούς τύπους.*

	N-άκρο	N+1	N+2	N+3	N+4	$N^1/_2$	$C^1/_2$	C-4	C-3	C-2	C-1	C-άκρο
T	18.7	8.6	1.5	1.9	0.5	6.0	6.8	6.4	2.8	2.8	6.3	4.7
D	19.6	8.9	10.4	7.7	9.4	6.2	3.9	7.7	2.5	6.4	8.6	9.1
P	2.2	17.5	1.5	1.5	0.0	0.8	0.4	0.3	0.0	0.0	0.0	1.3
E	2.5	2.8	20.1	21.3	4.1	6.9	5.2	2.7	2.0	15.0	3.8	0.5
Q	1.4	7.6	11.1	13.2	0.5	3.2	4.7	4.3	5.0	13.2	4.8	1.6
C	0.0	0.0	0.0	1.4	3.9	4.2	2.5	0.5	0.0	0.0	0.0	0.0
A	9.7	17.1	10.6	5.9	26.5	14.3	11.6	7.2	22.5	10.0	8.0	8.8
H	2.5	1.4	4.2	0.6	4.4	3.7	3.1	7.1	0.3	0.9	0.9	0.4
V	1.4	5.8	6.9	7.1	5.3	4.5	4.1	13.6	1.4	2.6	2.9	1.1

L	4.1	3.6	3.5	9.1	11.6	12.5	13.5	26.0	8.7	4.1	14.0	2.5
M	0.0	0.8	0.3	4.1	0.7	2.7	4.3	7.4	2.8	0.0	4.7	0.0
I	0.0	0.5	3.0	3.2	0.6	5.2	7.3	2.9	7.1	1.3	0.0	0.0
S	7.4	3.2	1.2	4.8	1.0	2.7	4.5	1.2	8.9	2.5	3.4	2.0
N	9.9	0.6	10.7	3.5	4.1	3.4	7.4	0.4	16.6	5.4	1.8	4.9
F	4.7	1.1	0.0	1.9	3.5	4.1	2.6	4.3	12.1	6.3	10.5	0.0
R	3.6	3.1	4.0	7.4	9.0	6.9	5.7	0.0	1.9	15.3	7.1	9.9
Y	3.6	1.6	0.0	0.6	0.5	1.3	2.4	0.8	2.7	0.0	8.8	0.0
K	7.4	9.6	9.1	1.4	12.5	8.8	5.9	2.2	1.4	12.0	11.7	12.4
W	0.0	3.8	0.0	0.6	0.0	0.7	0.2	0.0	0.0	0.0	0.0	4.6
G	1.3	2.2	2.2	2.8	1.9	2.0	3.7	4.8	1.2	2.3	2.7	36.1

Πίνακας 5. *Σχετική προτίμηση των είκοσι αμινοξικών τύπων για αξονικές θέσεις*

	N-άκρο	N+1	N+2	N+3	N+4	$N^1/_2$	$C^1/_2$	C-4	C-3	C-2	C-1	C-άκρο
T	3.02	1.39	0.24	0.31	0.08	0.97	1.10	1.03	0.45	0.45	1.02	0.76
D	2.80	1.27	1.49	1.10	1.34	0.89	0.56	1.10	0.36	0.91	1.23	1.30
P	1.47	11.67	1.00	1.00	0.00	0.53	0.27	0.20	0.00	0.00	0.00	0.87
E	0.37	0.41	2.96	3.13	0.60	1.02	0.77	0.40	0.29	2.21	0.56	0.07
Q	0.27	1.46	2.14	2.54	0.10	0.62	0.90	0.83	0.96	2.54	0.92	0.31
C	0.00	0.00	0.00	0.88	2.44	2.63	1.56	0.31	0.00	0.00	0.00	0.00
A	0.79	1.39	0.86	0.48	2.15	1.16	0.94	0.59	1.83	0.81	0.65	0.72
H	0.83	0.47	1.40	0.20	1.47	1.23	1.03	2.37	0.10	0.30	0.30	0.13
V	0.30	1.26	1.50	1.54	1.15	0.98	0.89	2.96	0.30	0.57	0.63	0.24
L	0.37	0.33	0.32	0.83	1.06	1.14	1.23	2.36	0.79	0.37	1.27	0.23
M	0.00	0.28	0.10	1.41	0.24	0.93	1.48	2.55	0.97	0.00	1.62	0.00
I	0.00	0.13	0.79	0.84	0.16	1.37	1.92	0.76	1.87	0.34	0.00	0.00
S	1.90	0.82	0.31	1.23	0.26	0.69	1.15	0.31	2.28	0.64	0.87	0.51
N	1.77	0.11	1.91	0.63	0.73	0.61	1.32	0.07	2.96	0.96	0.32	0.88
F	1.12	0.26	0.00	0.45	0.83	0.98	0.62	1.02	2.88	1.50	2.50	0.00
R	0.60	0.52	0.67	1.23	1.50	1.15	0.95	0.00	0.32	2.55	1.18	1.65
Y	1.90	0.84	0.00	0.32	0.26	0.68	1.26	0.42	1.42	0.00	4.63	0.00
K	1.01	1.32	1.25	0.19	1.71	1.21	0.81	0.30	0.19	1.64	1.60	1.70
W	0.00	4.75	0.00	0.75	0.00	0.88	0.25	0.00	0.00	0.00	0.00	5.75
G	0.30	0.50	0.50	0.64	0.43	0.46	0.84	1.09	0.27	0.52	0.61	8.21

Για τους "κυρίαρχους" αυτούς συνδυασμούς αξονικών θέσεων / αμινοξικών τύπων, ελέγχθηκε αν τροποποιούνται οι ακτινικές προτιμήσεις των τελευταίων. Τα N- και C-άκρα συνήθως δεν αντιστοιχούν σε κάποια από τις θέσεις a-g οπότε δεν τίθεται καν θέμα. Η προλίνη, όταν εμφανίζεται σε θέση N+1, καταλαμβάνει ακτινικά -όπως και στον Πίνακα A.I.3 - μια από τις θέσεις b, f, ή e. Η αλανίνη, που την συνοδεύει, δείχνει την παρατηρηθείσα και στα προηγούμενα ομοιόμορφη κατανομή στις θέσεις a-g. Το γλουταμικό κρατά στις θέσεις N+2, N+3 και C-2 την υψηλή προτίμησή του για θέση c, ενώ η γλουταμίνη δείχνει την ίδια ομοιόμορφη κατανομή σε ακτινικές θέσεις που έδειξε και στον Πίνακα A.I.3 . Η θέση N+4, αν και εμφανίζεται στο δείγμα μας στα 4/5 των περιπτώσεων σε υδρόφοβη θέση (a, d, e, g) όπως

και στην ανάλυση των Richardson και Richardson [1988], φιλοξενεί εξίσου και υδρόφιλα κατάλοιπα, ενώ δεν φαίνεται να υπάρχει κάποιος λόγος που η θέση C-4 καταλαμβάνεται σε υψηλά ποσοστά από υδρόφοβα (κυρίως λευκίνη και βαλίνη). Τέλος, η θέση C-1, για την οποία δεν ξεχωρίζει κάποιος κυρίαρχος αμινοξικός τύπος, και η θέση C-3 με τα υψηλό ποσοστό αλανίνης, σερίνης και ασπαραγίνης, μοιράζονται ακτινικά σε δύο θέσεις (Πίνακας 3).

Συμπεράσματα-Συζήτηση

Σε αντίθεση με τη έννοια της β-στροφής, που αποτέλεσε αντικείμενο μελέτης από τόσο νωρίς, ώστε η δημοσιευμένη το 1977 ανάλυση των Chou και Fasman να είναι μια πρώτη "ώριμη" ανάλυση σε 29 δομές πρωτεϊνών, τα συνδετικά τμήματα μεταξύ των α-ελίκων έπρεπε να περιμένουν τις αναλύσεις του Efimov, που -για δύο αντι-παράλληλα πακεταρισμένες α-έλικες, και σε κάπως ολοκληρωμένη μορφή- δημοσιεύτηκαν το 1991. Ίσως αιτία γι'αυτό να είναι η ανομοιομορφία που παρουσιάζουν, αφού -όπως δείχτηκε και στο παρόν κεφάλαιο- ***καθώς το μήκος τους αυξάνει, οδηγεί σε συμβατότητα με διαφορετικές -και όλο και περισσότερες- διαμορφώσεις***. Μπορεί κανείς να ισχυριστεί ότι η ανομοιομορφία για τις β-στροφές είναι μεγαλύτερη, αφού μπορούν να βρίσκονται ανάμεσα σε α-έλικες και β-κλώνους σε οποιοδήποτε συνδυασμο ή/και να αποτελούν μέρος μεγαλύτερων συνδετικών τμημάτων, ενώ τα συνδετικά τμήματα, με τα οποία συγκρίνουμε, είναι πάντα μεταξύ α-ελίκων. Όμως, *μια β-στροφή έχει πάντα μήκος τέσσερα αμινοξικά κατάλοιπα*, με συγκεκριμένη (την αυστηρή) απαίτηση να απέχουν τα δύο ακραία α-άτομα άνθρακα λιγότερο από 7Ε, και -σε ένα πιο αυστηρό ορισμό, που δεν ακολούθησαν οι Chou και Fasman [1977]- να υπάρχει δεσμός υδρογόνου ανάμεσα στο καρβονυλικό οξυγόνο της πρώτης και το άζωτο της τέταρτης θέσης.

Με τα παραπάνω υπ'όψη, στην εργασία του Efimov [1991] συνοψίζονται διαμορφώσεις εισόδων σε και εξόδων από α-έλικες που επιτρέπονται στερεοδιαταξικά, μαζί με κάποιες "θέσεις-κλειδιά" για κατάλοιπα υδρόφοβα ή γλυκίνης, ενώ σαν πλήρη συνδετικά τμήματα παρουσιάζονται μόνο σε θεωρητικό επίπεδο (χωρίς να έχει γίνει μια συστηματική καταγραφή με βάση πχ. μια τράπεζα δεδομένων) και με λίγα παραδείγματα, κάποια που αντιστοιχούν πάντα σε βραχέα τμήματα του παρόντος κεφαλαίου (συμπεριλαμβανομένου και του εξαιρετικά βραχέος που περιγράφεται για το κυτόχρωμα c'). Έκτοτε, άλλες εργασίες έχουν εστιάσει σε υποπεριπτώσεις άκρων ελίκων, που υπερκαλύπτονται εννοιολογικά από τη γενικότερη -και προηγειθείσα- ανάλυση των Richardson και Richardson [1988], ή σε συγκεκριμένα συνδετικά τμήματα [πχ Vlassi et al, 1994]. Πάντως, και για όσα συνδετικά τμήματα της παρούσης είναι μακρύτερα, από εκείνα που περιγράφει ο Efimov [1991], ισχύουν, για την εξοδό τους από την αμινοτελική και την είσοδό τους στην καρβοξυτελική α-έλικα, τα όσα αναφέρονται εκεί ότι αναμένονται θεωρητικά για την διαμόρφωσή τους, με τη διαφορά ότι, οι ακριβείς θέσεις, που ευνοούν υδρόφοβα κατάλοιπα, επηρεάζονται σε μεγάλο βαθμό από την έννοια στροφής του συνδετικού τμήματος (δεξιό-/αριστερό-στροφη, Εικόνα 1· πχ. τα όσα αναφέρθηκαν σε συνδυασμό με την Εικόνα 2). ***Όμως, ο αυξημένος αριθμός συνδυασμών για τμήματα με κάπως μεγαλύτερο μήκος σε καμία περίπτωση δεν δικαιολογεί τον όρο "τυχαία δομή"*** (ειδικά αφού μετά κάποιο μήκος αποκτούν δική τους διακριτή δομή -πχ. α-έλικας).

Ένα από τα σημεία που ίσως προβληματίσει τον αναγνώστη αυτού του κεφαλαίου, είναι η διαφορά στον ορισμό των άκρων των α-ελίκων, σε σχέση με το Κεφ. Α.Ι. Ο πολύ απλός

λόγος, που συμβαίνει αυτό, βρίσκεται στα διαφορετικά ερωτήματα των δύο κεφαλαίων. Πιο συγκεκριμένα, στο προηγούμενο κεφάλαιο, το ερώτημα ήταν αν οι θέσεις, *που συμμετέχουν στο δεμάτι*, εμφανίζονται κατειλημμένες σε μεγαλύτερα ποσοστά από κάποιους αμινοξικούς τύπους (σε σχέση με τους υπόλοιπους). ***Δεν ήταν καν αναγκαίο να συμπίπτουν τα άκρα των τμημάτων του Πίνακα Α.Ι.1 με τα άκρα των α-ελίκων.*** Στο τρέχον κεφάλαιο, τα ερωτήματα είναι διαφορετικά: (α) Τι είδους συνδετικά τμήματα (από πλευράς διαμόρφωσης, σε όρους διέδρων γωνιών της κύριας αλυσίδας) συνδέουν τις α-έλικες ενός δεματιού; (β) Τα άκρα των α-ελίκων, όπου απολήγουν αυτά τα συνδετικά τμήματα, έχουν συγκεκριμένες προτιμήσεις για κάποιες θέσεις a-g, ή/και για κάποιους αμινοξικούς τύπους; (γ) *Εαν και όποτε* κάποιες από τις τέσσερις θέσεις, αμέσως μέσα από τα άκρα των α-ελίκων, ανήκει *και* στο δεμάτι, αλλάζουν οι κατανομές του προηγουμένου κεφαλαίου; ***Ο Πίνακας Α.ΙΙ.1, όπως φανερά απαιτείται από τα αντίστοιχα ερωτήματα, αναφέρει τα άκρα των α-ελίκων, που -σημειωτέον- σε ορισμένες περιπτώσεις*** (που προκύπτουν με απλή διασταύρωση τους Πίνακες Α.Ι.1 και Α.ΙΙ.1, όπως πχ. το Ν-άκρο της τέταρτης α-έλικας της πρωτεΐνης του καλύμματος του ιού TMV) ***είναι εντελώς εκτός των ορίων του δεματιού.*** Εδώ να σημειωθεί ότι, οι Richardson και Richardson [1988] που ανέλυσαν τις τέσσερις ακραίες θέσεις όλων των τότε γνωστών α-ελίκων (από πλευράς τρισδιάστατης δομής εννοείται -συν δύο θέσεις έξω από τα άκρα), επέλεξαν εμπειρικά (ανάμεσα και από άλλους) τον ορισμό που χρησιμοποιείται και εδώ, με κριτήριο τη σαφήνεια τα αποτελέσματα που προέκυπταν. Εξάλλου, η επιλογή -στα πλαίσια της παρούσας- να μην διερευνηθούν θέσεις έξω από τα άκρα (αφού συχνά η πρώτη θέση έξω από το C-άκρο μιας α-έλικας είναι το Ν-άκρο της επόμενης), όχι μόνο δικαιολογείται από το μικρό μήκος των συνδετικών τμημάτων, αλλά είναι και συνεπής με την παρατήρηση των ιδίων ερευνητών ότι δεν υπήρχε κανένα σήμα για προτιμήσεις εκτός των ορίων που ορίζουν τα άκρα των ελίκων.

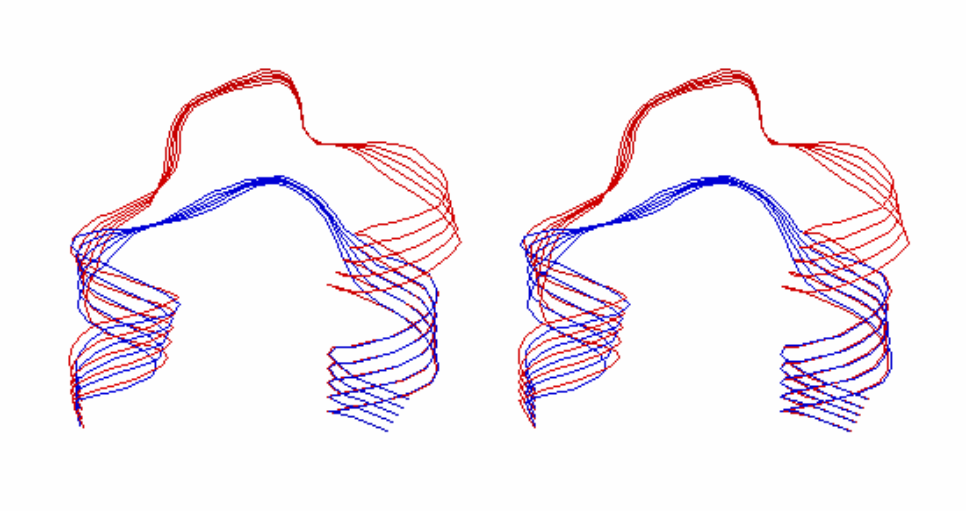
Όσον αφορά το είδος των συνδετικών τμημάτων μεταξύ των α-ελίκων (ερώτημα α), μπορούν να λεχθούν (κατά κατηγορία) τα εξής: (1) Για τα βραχέα συνδετικά τμήματα, από τις λίγες διαμορφώσεις που θεωρητικά αναμένονται [Efimov, 1991] ξεχωρίζει μία, η οποία και περιγράφτηκε σε λεπτομέρεια. (2) Για τα μεσαίου μεγέθους συνδετικά τμήματα, δεν παρατηρήθηκε κάποια αντίστοιχη κυριαρχία, καθώς οι αριθμοί πιθανών διαμορφώσεων αυξάνουν. Συχνά όμως, επιμέρους τμήματά τους ακολουθούν διαμορφώσεις σαν αυτές των βραχέων συνδετικών τμημάτων. (3) Τα μακρά συνδετικά τμήματα, όπως έχει ήδη αναφερθεί, έχουν ειδικές κατά οικογένεια διαμορφώσεις.

Ποιοι παράγοντες καθορίζουν το μήκος ενός συνδετικού τμήματος; Επειδή οι διαφορές στις αποστάσεις μεταξύ γειτονικών ελίκων (που συνδέονται μ'αυτό) θα μπορούσαν να είναι σημαντικές, αναζητήθηκαν οι (ακτινικά διακριτές) θέσεις, που πακετάρονται ανάμεσά τους. Προέκυψε ότι πρόκειται για τις θέσεις a και g: η σύσταση αυτών των θέσεων ενδέχεται να καθορίζει την απόσταση μεταξύ των ελίκων, θέτοντας ένα ελάχιστο απαιτούμενο μήκος.

Πράγματι, οι αιμερυθρίνες που έχουν αυξημένο ποσοστό λευκίνης (σε βάρος της αλανίνης) στις θέσεις a, έχουν και μεγαλύτερες αποστάσεις μεταξύ των αξόνων των α-ελίκων τους (σε σχέση για παράδειγμα με την πρωτεΐνη ROP), πράγμα που ίσως δικαιολογεί τα μακρύτερα συνδετικά τμήματα. Επίσης, περισσότερα κατάλοιπα μπορεί να απαιτούνται για την ίδια απόσταση μεταξύ ελίκων, εαν συμπεριλαμβάνεται μια παράλληλη είσοδος ή έξοδος. Περίσσεια καταλοίπων (από το απαιτούμενο ελάχιστο) οδηγεί, άλλοτε σε μικρές έλικες (όπως αναφέρθηκε για την λυσοζύμη) και άλλοτε σε επιμήκυνση κατά ένα μέρος της (των) έλικας(-ων) του δεματιού ή/και του συνδετικού τμήματος (όπως πχ. μεταξύ 3ης και 4ης έλικας στην οικογένεια των αιμερυθρινών, όπου οι μυοαιμερυθρίνες έχουν 5 επιπλέον κατάλοιπα -Εικόνα 5).

> Κατά τη διάρκεια της εργασίας του Κεφ. Α.Ι, οι θέσεις a και d θεωρήθηκαν εξίσου κρυμμένες στο εσωτερικό του δεματιού. Η βάση γι'αυτό βρίσκεται σε ένα υπερ-απλοϊκό σχήμα, που χρησιμοποιούταν συστηματικά στη βιβλιογραφία τότε (πχ Banner et al [1987] στην περιγραφή της πρωτεΐνης ROP -αλλά και αργότερα, όπως στις εργασίες του DeGrado σε "ελάχιστα" δεμάτια πχ DeGrado et al, [1989 αλλά και 1999!]), στο οποίο οι θέσεις a και d φαίνονται πακεταρισμένες με τον ίδιο τρόπο, δεξιά και αριστερά από τον κύριο άξονα του δεματιού (τον παράλληλο προς τον άξονα των ελίκων). Μάλιστα, όπως αναφέρθηκε στο Κεφ. Α.Ι, η κατάταξη των θέσεων στις κατηγορίες a-g έγινε οπτικά, γιατί δεν υπήρχαν άλλα, αντικειμενικά κριτήρια. Ανατρέχοντας όμως στην Εικόνα 7/Γεν.Εισαγωγή, διαπιστώνει κανείς ότι μόνο η θέση d δείχνει προς τον κύριο άξονα του δεματιού, ενώ η θέση a πακετάρεται προς μια γειτονική έλικα· μάλιστα, η θέση g που προηγείται στην αλληλουχία πακετάρεται προς την "άλλη" γειτονική έλικα. Αυτά τα χαρακτηριστικά έκαναν πιο σαφή τον ορισμό των θέσεων: η θέση d δείχνει τον κύριο άξονα του δεματιού, ενώ δύο διαδοχικές θέσεις που δείχνουν από μια (διαφορετική) γειτονική έλικα η καθεμία, είναι οι θέσεις g και a.

Όσον αφορά τις ακτινικές προτιμήσεις των ελικικών άκρων (ερώτημα β), οι α-έλικες ξεκινούν κατά προτίμηση σε συγκεκριμένες ακτινικές θέσεις, τις ίδιες όμως για μικρά και μεγάλα συνδ. τμήματα. Όμως, για τα μικρά μπορεί να αποτελούν απαίτηση: η διαμόρφωση, που υιοθετεί το βραχύ συνδετικό τμήμα, που περιγράφτηκε σε λεπτομέρεια στα αποτελέσματα, ίσως αποτελεί το μοναδικό τρόπο σύνδεσης δύο α-ελίκων με δύο μόνο κατάλοιπα (ανάμεσα στα τελευταία κατάλοιπα των ελίκων που είναι ακόμη σε ελικοειδή διαμόρφωση)· εφόσον απολήγει σε συγκεκριμένες (ακτινικές) θέσεις -που εξαρτώνται *μόνο* από την έννοια στροφής (δεξιά/αριστερά -Εικόνα 2)- οι θέσεις αυτές "επιβάλλονται" στα σχετικά N/C-άκρα.



Εικ. 5*. Στερεοσκοπική σχηματική αναπαράσταση του μέσου μήκους συνδετικού τμήματος μεταξύ των α-ελίκων υπ'αριθ. 3 και 4 στις αιμερυθρίνες (βραχύτερο-μπλε) και την μυοαιμερυθρίνη (κόκκινο), μετά από αλληλυπέρθεση των τμημάτων των συνδεομένων ελίκων, που είναι κοινά στα δύο μόρια. Η αμινοτελική α-έλικα "3" είναι αριστερά και η καρβοξυτελική α-έλικα "4" δεξιά, ο δε υδρόφοβος πυρήνας "πίσω" από το επίπεδο του χαρτιού. Η μυοαιμερυθρίνη έχει, σε αυτό το σημείο, μια εισοχή πέντε καταλοίπων σε σχέση με τις αιμερυθρίνες. Φαίνεται, πως κάποια από αυτά έχουν διευθετηθεί σε επιμήκυνση της καρβοξυτελικής έλικας (άνω δεξιά), και κάποια σε επιμήκυνση του συνδ. τμήματος. (Όσον αφορά την αιμερυθρίνη, πρόκειται για το συνδετικό τμήμα από την Εικόνα 3.)*

Στην αντίστοιχη καταμέτρηση συμμετέχει μόνο ο αντιπρόσωπος με γνωστή δομή, δηλαδή πρόκειται για 28 N- και 28 C-άκρα, που κατανέμονται σε ένα πίνακα (5+5) σειρές x 8 στήλες. Κι όμως, τα δεδομένα επαρκούν για να εξαχθούν ασφαλή συμπεράσματα. Αφήνοντας στην άκρη τις μικρές ανωμαλίες, που αναφέρθηκαν για τις ακραίες στροφές κάποιων ελίκων, ας υποτεθεί ότι δίνονται 7 οικογένειες, με έναν αντιπρόσωπο με γνωστή δομή η καθεμία, και ότι τα N- και C-άκρα τους τερματίζουν "ομοιόμορφα" στα όρια του δεματιού (δηλαδή δεν υπάρχουν περιπτώσεις όπου πχ. ολόκληρη η ακραία στροφή να είναι εκτός των ορίων του δεματιού). Ας υποτεθεί, επίσης, ότι σε ένα γενικό πληθυσμό από δεμάτια, απ'όπου επιλέχθηκε το δείγμα μας, δεν υπάρχουν προτιμήσεις για κάποιες (από τις επτά ακτινικά διακριτές) θέσεις. Τότε, στο γενικό πληθυσμό, 1/7 των N-άκρων αρχίζουν σε θέση a, 1/7 σε θέση b, κοκ. Αφού επιλέχθηκαν 7 οικογένειες -και άρα 28 N-άκρα- έπεται ότι 4 N-άκρα αρχίζουν σε θέση a, 4 σε θέση b, κοκ. Η πιθανότητα να βρεθούν ακριβώς $\boldsymbol{k}$ N-άκρα (από ένα σύνολο $\boldsymbol{n} >= \boldsymbol{k}$) στην ακτινικά ορισμένη θέση $\boldsymbol{j}$, για την οποία η *a priori* πιθανότητα είναι $\boldsymbol{p_j}$ (ίση με 1/7 εδώ), ακολουθεί την δυωνυμική κατανομή και είναι

$$\boldsymbol{P(j,k,n:p_j)}=(\boldsymbol{n} \parallel \boldsymbol{k})\cdot\boldsymbol{p_j^k}\cdot(1-\boldsymbol{p_j})^{\boldsymbol{n-k}}, \text{ όπου } (\boldsymbol{n} \parallel \boldsymbol{k})=[\boldsymbol{n}! / (\boldsymbol{k}! \cdot (\boldsymbol{n-k})!)]$$

με μέσο όρο $\boldsymbol{n\cdot p_j}$=28/7=4 και τυπική απόκλιση $\boldsymbol{s}=\sqrt{[\boldsymbol{n\cdot p_j}\cdot(1-\boldsymbol{p_j})]}= \sqrt{[28*(1/7)*(6/7)]}=1.85$.

Σύμφωνα με το σύνηθες όριο απόρριψης για μια υπόθεση (δηλαδή όταν έχει λιγότερο από 5% πιθανότητα να ισχύει), συνδυασμοί με εμφάνιση είτε μηδενική είτε μεγαλύτερη από 7/28 (δηλαδή 25%) είναι σημαντικοί (δηλαδή: αν μια θέση N+1…4 κατανέμεται εξίσου στις θέσεις a-g στον γενικό πληθυσμό, τότε το ποσοστό, που αυτή η θέση N+1…4 εμφανίζεται σαν *μια* θέση a-g, είναι μεγαλύτερο από 25% μόνο σε (λιγότερο από) 5% τυχαίων δειγμάτων με μέγεθος 7 οικογένειες), ενώ το 8/28(=28%) είναι σημαντικό και στο όριο -περίπου- του 1% . Φυσικά η εκτίμηση αυτή αφορά μια αξονική σε συνδυασμό με μια ακτινική θέση· για το σύνολο των επτά πλαισίων αντιστοίχισης του Πίνακα 3, ισχύει η πολυωνυμική, και η φυσικά η πιθανότητα να ανήκει συνολικά σε μια ομοιόμορφη κατανομή είναι πολύ μικρότερη απ'ότι για τα επιμέρους πλαίσια (3/7 πλαίσια αντιστοίχισης είναι σημαντικά στο 5%, από τα οποία το ένα στο 1%). Για τα βραχέα συνδετικά τμήματα, η άποψη ότι δεν πρόκειται για τυχαίο γεγονός, ενισχύεται επιπλέον από το μικρό αριθμό, καλά καθορισμένων, προσβάσιμων διαμορφώσεων.

Εξάλλου, στις περισσότερες από τις αξονικά ορισμένες θέσεις, ορισμένοι αμινοξικοί τύποι απαντούν συχνότερα από άλλους. Η σημαντικότητα των υψηλών (και χαμηλών) ποσοστών για κάποιους συνδυασμούς αμινοξικών τύπων / αξονικά ορισμένων θέσεων στον Πίνακα 4, στατιστικά αναλύεται με τρόπο ανάλογο με τα παραπάνω, ενώ συχνά ενισχύεται και από το γεγονός ότι είναι αμινοξικοί τύποι με παρόμοια χαρακτηριστικά που παρουσιάζουν ταυτόχρονα ένα υψηλό (ή χαμηλό) ποσοστό, προφανώς αντανακλώντας στερεοδιαταξικές και φυσικοχημικές ανάγκες των συγκεκριμένων θέσεων που ερμηνεύονται απλά. Οι DeGrado et al [1999] συνοψίζουν σε ένα άρθρο ανασκόπησης, ανάμεσα σε άλλα, αποτελέσματα από εργασίες (δημοσιευμένες πολύ μετά το πέρας της παρούσας) που δείχνουν ότι ένα N-άκρο, με χαρακτηριστικά σαν αυτά που περιγράφονται εδώ (δυνατότητα δεσμού υδρογόνου ή υδρόφοβης αλληλεπίδρασης με κατάλοιπα από επόμενη στροφή), μπορεί να σταθεροποιεί μια α-έλικα μέχρι και 2 Kcal/mol, ενώ ένα C-άκρο σε διαμόρφωση α_L , που δεν την σταθεροποιεί, αλλά μάλλον είναι απαραίτητο να είναι σε αυτή τη διαμόρφωση για στερεοδιαταξικούς λόγους, προτιμά γλυκίνη. Εξάλλου, ένα κατάλοιπο προλίνης, ενώ στο μέσο μιας α-έλικας είναι "προβληματικό[23]", στη θέση N+1, όπου δεν υπάρχει στερεοδιαταξικό πρόβλημα από "επόμενη στροφή της έλικας", γίνεται αιτία (1) να μην μένει εκτεθειμένος στο διαλύτη δεσμός υδρογόνου μη-ικανοποιημένος από την αλυσίδα και (2) να μειώνεται η απώλεια εντροπίας που προκαλεί η ακινητοποίηση λόγω διπλώματος (η προλίνη, στην κατάσταση αποδιάταξης, έχει

[23] Πιο συγκεκριμένα, προκαλεί "τσακίσμα" της α-έλικας. Προφανώς όμως, η επιλογή (για κάποιες πρωτεΐνες, πχ. κυτοχρώματα στην παρούσα) ενός τέτοιου χαρακτηριστικού εξελικτικά, σημαίνει ότι είναι μάλλον πλεονεκτικό, παρά προβληματικό -όθεν και τα εισαγωγικά.

μικρότερη ελευθερία περιστροφής, σε σχέση με άλλους αμινοξικούς τύπους). Κοινό τόπο αποτελεί επίσης η συχνή εμφάνιση Asp, Asn, Glu και Gln στην πρώτη στροφή στις α-έλικες.

Πάντως, στο δεμάτι δεν παρατηρείται το C-άκρο σε διαμόρφωση α_L στο υψηλό ποσοστό που αναφέρουν οι DeGrado et al [1999] (75% από 261 περιπτώσεις που "εξέτασαν οι ίδιοι", χωρίς άλλη αναφορά σε πρωτότυπη εργασία) -αλλιώς η γλυκίνη ίσως έχαιρε ποσοστού πολύ ανώτερου του 36% . Επίσης, αναφέρουν ότι τα πρότυπα (έλικα)-α_L-β και β-(έλικα), με το κατάλοιπο σε διαμόρφωση β να αποτελεί το N-άκρο της δεύτερης α-έλικας, είναι συχνά και -άρα- αναμένεται να εμφανίζεται συχνά το πρότυπο (έλικα)-α_L-β-β-(έλικα). Όμως, ούτε κάτι τέτοιο προκύπτει από τους πίνακες του κεφαλαίου. Αντίθετα, το (έλικα)-α_L-β/δ-(έλικα) που δίνουν σαν δευτερεύουσα εναλλακτική, είναι ο κύριος τύπος για τα βραχέα συνδετικά τμήματα.

Τέλος, ***όσον αφορά την επίδραση των παραπάνω στις κατανομές του δεματιού σε ακραίες θέσεις των ελίκων (ερώτημα γ),*** μπορούμε να πούμε ότι: (1) Τα N- και C-άκρα με τις πλέον ιδιότυπες κατανομές, συνήθως δεν μετέχουν στο δεμάτι, (2) Στις υπόλοιπες θέσεις, και ένα ομαλότερο ισοζύγιο υπάρχει για τις περισσότερες, και οι αμινοξικοί τύποι που παρατηρούνται -χωρίς να παύουν να είναι σύμφωνοι με τα αποτελέσματα των Richardson και Richardson [1988]- είναι γενικά σύμφωνοι με την ακτινική θέση που καταλαμβάνουν, ακόμα και (κυρίως εκεί) για τις περιπτώσεις που κάποια ποσοστά εμφάνισης ξεφεύγουν από την έννοια του "ομαλού ισοζυγίου" (πχ το γλουταμικό οξύ που κυριαρχεί στη θέση N+2 διατηρεί την προτίμησή του για θέση c). Πράγματι, οι περιπτώσεις, με αφορμή την παρατήρηση των οποίων, ξεκίνησε η ανάλυση των συνδετικών τμημάτων και των ακραίων στροφών, είναι μάλλον αποσπασματικές και συγκεκριμένες: αρωματικά σε θέσεις a/d στα όρια ελίκων και συνδετικών τμημάτων, και γλυκίνη στο C-άκρο -συχνά εκτός ορίων δεματιού-, ενώ η προλίνη στη θέση N+1, καθώς και 2-3 αμινοξικοί τύποι, που ξεχωρίζουν για τις υπόλοιπες θέσεις των ακραίων στροφών, δεν μεταβάλλουν την ακτινική τους προτίμηση. Το αντίστροφο πρέπει να αναμένεται ότι δεν ισχύει: η σύσταση των θέσεων a-g ίσως μεταβάλλεται όταν βρίσκονται σε ακραίες στροφές, αλλά με την έννοια της περαιτέρω έντασης ή καταστολής χαρακτηριστικών που ήδη έχουν (και όχι με την έννοια ότι πχ. ξαφνικά οι θέσεις f θα γεμίσουν υδρόφοβα κατάλοιπα), αλλά στα πλαίσια που αφήνει το γεγονός ότι οι (όσες) θέσεις από τις ακραίες στροφές των α-ελίκων συμμετέχουν στο δεμάτι, αποτελούν ήδη σημαντικό ποσοστό των θέσεων που αναλύθηκαν στο Κεφάλαιο A.I (1/4 εως 2/3, ανάλογα με την οικογένεια). Στο ίδιο παραπάνω παράδειγμα, μια ματιά στους Πίνακες 4,5 αρκεί για να φανεί ότι μια θέση c, σαν θέση N+2 θα πρέπει να αναμένεται πιο συχνά κατειλημένη από γλουταμικό απ'ότι μια θέση c γενικώτερα και περισσότερο από μια θέση c σε "εσωτερική" στροφή της έλικας. Όμως, τα δεδομένα δεν επαρκούν για μια ανάλυση σε τόσες πολλές (7x10=70 τουλάχιστον) κατηγορίες (x20 αμινοξικούς τύπους!). Εαν τα δεδομένα του παρόντος κεφαλαίου μπορούν να χρησιμο-

ποιηθούν σε συνδυασμό με τα δεδομένα των ακτινικών κατανομών για να αυξήσουν σε επίπεδο πρόβλεψης την ακρίβεια προσδιορισμού των άκρων των ελίκων, αυτό μένει να δειχτεί.

Μένει τέλος ένα θέμα, τόσο απρόσιτο να λυθεί προς το παρόν, όσο και το ερώτημα αν προηγήθηκε το αυγό ή η κότα: οι α-έλικες, γενικότερα, έχουν προτιμήσεις για συγκεκριμένους αμινοξικούς τύπους στις θέσεις των αμινοτελικών και καρβοξυτελικών στροφών. Αυτοί οι αμινοξικοί τύποι, με τη σειρά τους, έχουν προτιμήσεις για συγκεκριμένες ακτινικά διακριτές θέσεις. Αυτές οι θέσεις, με τη σειρά τους, είναι που εξυπηρετούν καλύτερα σαν απολήξεις συνδετικών τμημάτων (συχνά με ελάχιστο μήκος!). Μήπως, κάποια από αυτές τις ανάγκες, σαν λιγότερο ελαστική, επιλέγει (και επιβάλλει) τις λύσεις με τις οποίες θα πρέπει να ικανοποιηθούν οι υπόλοιπες; Και ποια είναι η ανάγκη αυτή; Ή μήπως, μέσα από διάφορους (θεωρητικούς, πιθανούς, αλλά όχι παρατηρημένους) συνδυασμούς, έχουν εξελικτικά επιλεγεί "συνολικά" κάποιες λύσεις; Η συζήτηση επιστρέφει στο θέμα των καλών συνταιριασμάτων των διαφορετικών απαιτήσεων του δεματιού (το οποίο θέμα θα μείνει στη σφαίρα των υποθέσεων), σε μια εντονότερη έκφανση του, κατά την παρουσίαση της δημιουργίας ενός μοντέλλου με βάση τους πίνακες των Κεφαλαίων Α.Ι και Α.ΙΙ.

> Πάντως, εκ των υστέρων, όλα ακούγονται λογικά. Η άλλη εναλλακτική θα ήταν να δοθεί προτεραιότητα σε κάποια από τις ανάγκες κατασκευής του δεματιού χωρίς να ικανοποιηθούν άλλες απαιτήσεις του, τοποθετώντας έτσι πχ. αμινοξικά κατάλοιπα σε θέσεις όπου δεν ταιριάζουν στερεοδιαταξικά, και άρα αποσταθεροποιώντας το πρωτεϊνικό μόριο.

Προσάρτημα: Υπολογισμός τοπικών αξόνων

Όταν δύο συστήματα συντεταγμένων δεν σχετίζονται μόνο με απλή μετατόπιση, αλλά και περιστροφή γύρω από κάποιο άξονα, τότε το μοναδιαίο διάνυσμα του άξονα περιστροφής έχει τις ίδιες συντεταγμένες στους ισοδύναμους άξονες συντεταγμένων των δύο συστημάτων: η προβολή του στον άξονα x του ενός συστήματος είναι ίση με την προβολή του στον άξονα x του άλλου κοκ. Αυτή η σχέση ορίζει μια οικογένεια ευθειών παραλλήλων προς ένα (ελεύθερο) μοναδιαίο διάνυσμα, έστω **P** με προβολές P_x, P_y και P_z (σε εξωτερικό σύστημα συντεταγμένων πχ του παρατηρητή). Από την οικογένεια αυτή, σαν (σταθερός) άξονας περιστροφής μπορεί να επιλεγεί εκείνος που διέρχεται από σημείο τέτοιο ώστε (α) κάποιο σημείο αναφοράς κοινό στα δύο συστήματα να προβάλλει κάθετα στον άξονα και (β) οι κάθετες προβολές των αρχών των δύο συστημάτων επί του άξονα περιστροφής να έχουν τις ίδιες εσωτερικές συντεταγμένες (στο (τοπικό) σύστημά τους δηλαδή): ότι συντεταγμένη x έχει στο ένα σύστημα η προβολή της αρχής του συστήματος στον άξονα περιστροφής, την ίδια πρέπει να έχει η προβολή της αρχής του άλλου συστήματος πάνω στον άξονα περιστροφής, σε εκείνο το τοπικό σύστημα.

Η υπορουτίνα στο τέλος του κεφαλαίου κάνει ακριβώς αυτό. Στην αρχή, για κάθε αμινοξικό κατάλοιπο, όπου το άζωτο του πεπτιδικού δεσμού έχει συντεταγμένες (x_N, y_N, z_N), ο α-άνθρακας (x_{CA}, y_{CA}, z_{CA}) και ο καρβονυλικός άνθρακας (x_C, y_C, z_C), ορίζεται ένα (τοπικό) σύστημα συντεταγμένων με αρχή τον α-άνθρακα, τον άξονα x επί του δεσμού C_α-N, τον άξονα y επί του επιπέδου της (οξείας) γωνίας N-C_α-C και κάθετο στον άξονα x, και τον άξονα z κάθετο στους δύο προηγούμενους, και ώστε το σύστημα (x,y,z) να είναι δεξιόστροφο. Στα ακόλουθα, τα διανύσματα δίνονται μέσα σε παρενθέσεις, είτε με έντονα γράμματα όταν δίνονται με βάση τα δύο άκρα τους -πχ (**AB**) σημαίνει το διάνυσμα με άκρα τα σημεία A και B- είτε με απλά γράμματα όταν δίνονται με βάση τις συντεταγμένες τους -πχ, (AB_x, AB_y, AB_z) ή (x_{AB}, y_{AB}, z_{AB}) σημαίνουν το διάνυσμα AB με αντίστοιχες προβολές σε x, y και z.

Ορίζονται λοιπόν, τα διανύσματα $(\mathbf{C_\alpha N}) = (can_x=x_N-x_{CA},\ can_y=y_N-y_{CA},\ can_z=z_N-z_{CA})$ με μέτρο $|\mathbf{C_\alpha N}| = \sqrt{(can_x^2 + can_y^2 + can_z^2)}$, και μοναδιαίο διάνυσμα $(\mathbf{u_{can}}) = (\mathbf{C_\alpha N})/|\mathbf{C_\alpha N}|$, και $(\mathbf{C_\alpha C}) = (cac_x=x_C-x_{CA},\ cac_y=y_C-y_{CA},\ cac_z=z_C-z_{CA})$ με μέτρο $|\mathbf{C_\alpha C}| = \sqrt{(cac_x^2+ cac_y^2+ cac_z^2)}$, και μοναδιαίο διάνυσμα $(\mathbf{u_{cac}}) = (\mathbf{C_\alpha C})/|\mathbf{C_\alpha C}|$. Το μοναδιαίο διάνυσμα επί του άξονα y είναι τότε $(\mathbf{u_Y})=(\mathbf{Y})/|\mathbf{Y}|$, όπου το διάνυσμα $(\mathbf{Y}) = (\mathbf{C_\alpha C}) - (\mathbf{u_{can}})*[(\mathbf{C_\alpha N})\cdot(\mathbf{C_\alpha C})]/|\mathbf{C_\alpha N}|$ (ο αστερίσκος είναι ο κοινός πολλαπλασιασμός, ενώ η άνω τελεία δείχνει το εσωτερικό γινόμενο). Το μοναδιαίο επί του άξονα z είναι $(\mathbf{u_Z})=(\mathbf{Z})/|\mathbf{Z}|$, όπου το διάνυσμα $(\mathbf{Z}) = (\mathbf{C_\alpha N})x(\mathbf{Y})$ (το x εδώ είναι το εξωτερικό γινόμενο, που έχει ως αποτέλεσμα διάνυσμα).

Αν το ζητούμενο (ελεύθερο) μοναδιαίο διάνυσμα επί του τοπικού άξονα, έστω (**P**) έχει συντεταγμένες (P_x,P_y,P_z) σε σύστημα αναφοράς εξωτερικό ως προς τα αμινοξικά κατάλοιπα

(έστω εκείνο του παρατηρητή), μπορεί να αναλυθεί στο τοπικό σύστημα συντεταγμένων του καταλοίπου i σε προβολές λ_i, μ_i, και ν_i της μορφής:

$$\begin{aligned} \lambda_i &= uc\alpha n_x * P_x + uc\alpha n_y * P_y + uc\alpha n_z * P_z \\ \mu_i &= uY_x * P_x + uY_y * P_y + uY_z * P_z \\ \nu_i &= uZ_x * P_x + uZ_y * P_y + uZ_z * P_z \end{aligned} \qquad [1]$$

Φυσικά η ανάγκη να είναι οι προβολές λ, μ και ν ίδιες σε δύο διαδοχικά κατάλοιπα i και i-1, σημαίνει ότι

$$\begin{aligned} \{uc\alpha n_x[i]-uc\alpha n_x[i-1]\} * P_x + \{uc\alpha n_y[i]-uc\alpha n_y[i-1]\} * P_y + \{uc\alpha n_z[i]-uc\alpha n_z[i-1]\} * P_z = 0 \\ \{uY_x[i]-uY_x[i-1]\} * P_x + \{uY_y[i]-uY_y[i-1]\} * P_y + \{uY_z[i]-uY_z[i-1]\} * P_z = 0 \\ \{uZ_x[i]-uZ_x[i-1]\} * P_x + \{uZ_y[i]- uZ_y[i-1]\} * P_y + \{uZ_z[i]- uZ_z[i-1]\} * P_z = 0 \end{aligned} \qquad [2]$$

στο οποίο (ομογενές 3x3) σύστημα εξισώσεων οι μόνοι άγνωστοι είναι τα P_x, P_y και P_z. Το πρώτο μέρος της υπορουτίνας έχει τη λύση ακριβώς αυτού του συστήματος.

Στο δεύτερο μέρος, σαν σημείο αναφοράς κοινό στα δύο συστήματα επιλέγεται το μέσο του πεπτιδικού δεσμού που τα ενώνει· επιλέγεται δηλαδή ότι, στο σημείο από το οποίο διέρχεται ο άξονας περιστροφής, προβάλλει κάθετα το μέσο του πεπτιδικού δεσμού, που ενώνει τα δύο κατάλοιπα. Έτσι, αν M είναι το μέσο του πεπτιδικού δεσμού με συντεταγμένες (m_x, m_y, m_z), και x_{Ps}, y_{Ps} και z_{Ps} οι συντεταγμένες του ζητούμενου σημείου, ισχύει

$$[m_x-x_{Ps}]*P_x+[m_y-y_{Ps}]*P_y+[m_z-z_{Ps}]*P_z=0 \qquad [3]$$

όπου τα P_x, P_y και P_z υπολογίστηκαν στο πρώτο μέρος. Αυτή η συνθήκη καθετότητας, μαζί με την απαίτηση για ίχνη προβολών των α-ατόμων άνθρακα με ίσες εσωτερικές συντεταγμένες (που την επεξεργάζεται κανείς με τρόπο αντίστοιχο των σχέσεων [1] και [2]) οδηγούν σε ένα σύστημα τριών εξισώσεων με τρεις αγνώστους (3x3 κατά Cramer), του οποίου η λύση δίνεται στο δεύτερο μέρος της υπορουτίνας. Ο πίνακας A είναι ο πίνακας των υπο-οριζουσών του συστήματος, ενώ DETER είναι η ορίζουσα του συστήματος (που πρέπει να είναι μη-μηδενική). Οι υπολογισμοί αν και μακροσκελείς είναι απλοί και αποτελούν ύλη Λυκείου.

Η διαφορά της προσέγγισης αυτή από την μέθοδο των Sklenar και συνεργατών [1989] είναι ότι εκείνοι υπολογίζουν τον άξονα λαμβάνοντας υπ'όψη περισσότερα κατάλοιπα, οπότε τα παραπάνω συστήματα έχουν πολύ περισσότερες εξισώσεις από αγνώστους, και δεν είναι εγγυημένο ότι έχουν λύση (και συνήθως δεν έχουν) με την οποία να ικανοποιούνται επακριβώς όλες οι εξισώσεις. Έτσι λοιπόν ορίζουν μια συνάρτηση που δείχνει το μέγεθος της ασυμφωνίας αυτής, την οποία και προσπαθούν να ελαχιστοποιήσουν. Ίσως ο άξονας που υπολογίζεται έτσι να είναι ακριβέστερος υπό κάποιες συνθήκες, αλλά για περιπτώσεις όπως τα διαδοχικά κατάλοιπα μιας έλικας, η διαφορά στην πράξη (και ακόμη περισσότερο για τις ποιοτικές ανάγκες της δουλειάς αυτού του κεφαλαίου) είναι μικρή, και δικαιολογεί την (κατά πολύ

απλούστερη) προσέγγιση που περιγράφτηκε παραπάνω. Ας σημειωθεί όμως ότι και για την περίπτωσή τους, αυστηρά μιλώντας, δεν υπάρχει περιγραφή της συμπεριφοράς της συνάρτησής τους (πχ. πιθανά πολλαπλά τοπικά ελάχιστα), που σημαίνει ότι θεωρητικά δεν είναι βέβαιο ότι ο άξονας που υπολογίζεται έτσι είναι ο καλύτερος κάτω από όλες τις συνθήκες. Για παράδειγμα όντας υποχρεωτικά ευθύς επί πολλά διαδοχικά κατάλοιπα, δεν είναι βέβαιο ότι περιγράφει σωστά τις περιπτώσεις των έντονων τσακισμάτων που συχνά παρατηρούνται σε α-έλικες.

Συνάρτηση LoHAx - Δοθέντων των συντεταγμένων του αζώτου, του α-άνθρακα και του καρβονυλικού άνθρακα δύο διαδοχικών καταλοίπων, υπολογίζει τα συνημίτονα κατεύθυνσης (προβολές) και το σημείο εφαρμογής του μοναδιαίου διανύσματος του άξονα γύρω από τον οποίο πρέπει να περιστρέψει κανείς το ένα κατάλοιπο (για την ακρίβεια ένα σύστημα αναφοράς ορισμένο τοπικά με βάση τα άτομα του καταλοίπου) ώστε να πέσει πάνω στο άλλο.

```
int lohax (xn1,yn1,zn1,xca1,yca1,zca1,xco1,yco1,zco1,
           xn0,yn0,zn0,xca0,yca0,zca0,xco0,yco0,zco0,
           rPx,rPy,rPz,rxPs,ryPs,rzPs)

REAL4 xn1,yn1,zn1,xca1,yca1,zca1,xco1,yco1,zco1,
      xn0,yn0,zn0,xca0,yca0,zca0,xco0,yco0,zco0,
      *rPx,*rPy,*rPz,*rxPs,*ryPs,*rzPs; /* r stands for returned */

/* NOTE: In the following, index 0 --when present-- refers to residue i,
   whereas index 1 refers to i-1 !!! An appendix of 2 mean the square of
   the value. */
{
 int   MAXTRI, MAXDET;
 REAL4 Px,Py,Pz;
 REAL4 canx, canx1, cany, cany1, canz, canz1;
 REAL4 cacx, cacx1, cacy, cacy1, cacz, cacz1;
 REAL4 canl01, canl11, cacl01, cacl11;
 REAL4 ucanx, ucany, ucanz, ucanx1, ucany1, ucanz1, cancac, cancc1;
 REAL4 Yx,  Yy,  Yz,  Ylen,  Yux,  Yuy,  Yuz,  Zux,  Zuy,  Zuz;
 REAL4 Yx1, Yy1, Yz1, Ylen1, Yux1, Yuy1, Yuz1, Zux1, Zuy1, Zuz1;
 REAL4 D[4][4], TRI[4], axcoor[4], inprod;
 REAL4 mitx, mity, mitz, A[4][5], DETER, DX, DY, DZ;

/***ΜΕΡΟΣ Ι: Συνημίτονα Κατεύθυνσης ***/

/* Ορισμός διανυσμάτων κατά μήκος των δεσμών Ca-N και Ca-C... */
 canx =xn0-xca0;  canx1=xn1-xca1;  cany =yn0-yca0;
 cany1=yn1-yca1;  canz =zn0-zca0;  canz1=zn1-zca1;
 cacx =xco0-xca0; cacx1=xco1-xca1; cacy =yco0-yca0;
 cacy1=yco1-yca1; cacz =zco0-zca0; cacz1=zco1-zca1;

/* ...και υπολογισμός του μήκους τους */
 canl01=(REAL4)sqrt((double)( canx*canx + cany*cany + canz*canz ));
 canl11=(REAL4)sqrt((double)(canx1*canx1+cany1*cany1+canz1*canz1));
 cacl01=(REAL4)sqrt((double)( cacx*cacx + cacy*cacy + cacz*cacz ));
 cacl11=(REAL4)sqrt((double)(cacx1*cacx1+cacy1*cacy1+cacz1*cacz1));

/* Ορισμός τριγ-ορθοκανονικού συστήματος σύμφωνα με το κείμενο */
  ucanx=canx/canl01;   ucany=cany/canl01;   ucanz=canz/canl01;
 ucanx1=canx1/canl11; ucany1=cany1/canl11; ucanz1=canz1/canl11;

/* cancac είναι το εσωτερικό γινόμενο can.cac */
 cancac= canx*cacx + cany*cacy + canz*cacz;
 cancc1=canx1*cacx1+cany1*cacy1+canz1*cacz1;

 Yx=cacx-ucanx*(cancac/canl01);
 Yy=cacy-ucany*(cancac/canl01);
 Yz=cacz-ucanz*(cancac/canl01);
 Ylen=sqrt(Yx*Yx+Yy*Yy+Yz*Yz);
 Yux=Yx/Ylen; Yuy=Yy/Ylen; Yuz=Yz/Ylen;
 Yx1=cacx1-ucanx1*(cancc1/canl11);
 Yy1=cacy1-ucany1*(cancc1/canl11);
 Yz1=cacz1-ucanz1*(cancc1/canl11);
 Ylen1=(REAL4)sqrt((double)(Yx1*Yx1+Yy1*Yy1+Yz1*Yz1));
 Yux1=Yx1/Ylen1; Yuy1=Yy1/Ylen1; Yuz1=Yz1/Ylen1;

 Zux=ucany*Yuz-ucanz*Yuy;
 Zuy=ucanz*Yux-ucanx*Yuz;
 Zuz=ucanx*Yuy-ucany*Yux;
 Zux1=ucany1*Yuz1-ucanz1*Yuy1;
```

```
 Zuy1=ucanz1*Yux1-ucanx1*Yuz1;
 Zuz1=ucanx1*Yuy1-ucany1*Yux1;

/* Λύση για το ομογενές σύστημα 3x3. */
 D[1][1]=(Yuy-Yuy1)*(Zuz-Zuz1)-(Zuy-Zuy1)*(Yuz-Yuz1);
 D[1][2]=(Yux-Yux1)*(Zuz-Zuz1)-(Zux-Zux1)*(Yuz-Yuz1);
 D[1][3]=(Yux-Yux1)*(Zuy-Zuy1)-(Zux-Zux1)*(Yuy-Yuy1);
 D[2][1]=(ucany-ucany1)*(Zuz-Zuz1)-(Zuy-Zuy1)*(ucanz-ucanz1);
 D[2][2]=(ucanx-ucanx1)*(Zuz-Zuz1)-(Zux-Zux1)*(ucanz-ucanz1);
 D[2][3]=(ucanx-ucanx1)*(Zuy-Zuy1)-(Zux-Zux1)*(ucany-ucany1);
 D[3][1]=(ucany-ucany1)*(Yuz-Yuz1)-(Yuy-Yuy1)*(ucanz-ucanz1);
 D[3][2]=(ucanx-ucanx1)*(Yuz-Yuz1)-(Yux-Yux1)*(ucanz-ucanz1);
 D[3][3]=(ucanx-ucanx1)*(Yuy-Yuy1)-(Yux-Yux1)*(ucany-ucany1);
 TRI[1]=D[1][1]*D[1][1]+D[1][2]*D[1][2]+D[1][3]*D[1][3];
 TRI[2]=D[2][1]*D[2][1]+D[2][2]*D[2][2]+D[2][3]*D[2][3];
 TRI[3]=D[3][1]*D[3][1]+D[3][2]*D[3][2]+D[3][3]*D[3][3];

/* Επιλογή της ορίζουζας που το επιλύει ακριβέστερα... */
 MAXTRI=1; if (TRI[2] > TRI[1]) MAXTRI=2;
 if (TRI[3] > TRI[MAXTRI]) MAXTRI=3;
 if (TRI[MAXTRI] == (REAL4)0.0) return 0-1;

 MAXDET=1;
 if (fabs((double)D[MAXTRI][2]) > fabs((double)D[MAXTRI][1])) MAXDET=2;
 if (fabs((double)D[MAXTRI][3]) > fabs((double)D[MAXTRI][MAXDET])) MAXDET=3;

/* ...και επίλυση! */
 axcoor[MAXDET]=
  (REAL4)sqrt((double)(D[MAXTRI][MAXDET]*D[MAXTRI][MAXDET]/TRI[MAXTRI]));
 if (MAXDET == 3) {
  axcoor[1]=  axcoor[3]*D[MAXTRI][1]/D[MAXTRI][3];
  axcoor[2]=(-axcoor[3]*D[MAXTRI][2]/D[MAXTRI][3]);
 }/*endif*/
 if (MAXDET == 2) {
  axcoor[1]=(-axcoor[2]*D[MAXTRI][1]/D[MAXTRI][2]);
  axcoor[3]=(-axcoor[2]*D[MAXTRI][3]/D[MAXTRI][2]);
 }/*endif*/
 if (MAXDET == 1) {
  axcoor[2]=(-axcoor[1]*D[MAXTRI][2]/D[MAXTRI][1]);
  axcoor[3]=  axcoor[1]*D[MAXTRI][3]/D[MAXTRI][1];
 }/*endif*/

 inprod=axcoor[1]*(xca0-xca1)+axcoor[2]*(yca0-yca1)+axcoor[3]*(zca0-zca1);
 if (inprod < (REAL4)0.0)
  {axcoor[1]=(-axcoor[1]); axcoor[2]=(-axcoor[2]); axcoor[3]=(-axcoor[3]);}

 Px=axcoor[1]; *rPx=Px; Py=axcoor[2]; *rPy=Py; Pz=axcoor[3]; *rPz=Pz;

/***ΜΕΡΟΣ ΙΙ: Σημείο εφαρμογής ***/
/* ΣΗΜ: px, py και pz είναι τα συνημίτονα κατεύθυνσης για ένα μοναδιαίο
  διάνυσμα πάνω στον τοπικό άξονα. Οι απαιτήσεις που εξηγούνται στο κείμενο
  ορίζουν ένα σύστημα Cramer 3x3. Ακολουθεί η επίλυση. */

 mitx=(xco1+xn0)/(REAL4)2.0; mity=(yco1+yn0)/(REAL4)2.0;
 mitz=(zco1+zn0)/(REAL4)2.0;

 A[1][1]= (ucanx-ucanx1)*(Px*Px-1)+(ucany-ucany1)*Py*Px+(ucanz-ucanz1)*Pz*Px;
 A[1][2]= (ucanx-ucanx1)*Px*Py+(ucany-ucany1)*(Py*Py-1)+(ucanz-ucanz1)*Pz*Py;
 A[1][3]= (ucanx-ucanx1)*Px*Pz+(ucany-ucany1)*Py*Pz+(ucanz-ucanz1)*(Pz*Pz-1);
 A[1][4]= (ucanx*(Px*Px-1)+ucany*Py*Px+ucanz*Pz*Px)*xca0
         -(ucanx1*(Px*Px-1)+ucany1*Py*Px+ucanz1*Pz*Px)*xca1
         +(ucanx*Px*Py+ucany*(Py*Py-1)+ucanz*Pz*Py)*yca0
         -(ucanx1*Px*Py+ucany1*(Py*Py-1)+ucanz1*Pz*Py)*yca1
         +(ucanx*Px*Pz+ucany*Py*Pz+ucanz*(Pz*Pz-1))*zca0
         -(ucanx1*Px*Pz+ucany1*Py*Pz+ucanz1*(Pz*Pz-1))*zca1;

 A[2][1]= (Yux-Yux1)*(Px*Px-1)+(Yuy-Yuy1)*Py*Px+(Yuz-Yuz1)*Pz*Px;
 A[2][2]= (Yux-Yux1)*Px*Py+(Yuy-Yuy1)*(Py*Py-1)+(Yuz-Yuz1)*Pz*Py;
 A[2][3]= (Yux-Yux1)*Px*Pz+(Yuy-Yuy1)*Py*Pz+(Yuz-Yuz1)*(Pz*Pz-1);
 A[2][4]= (Yux*(Px*Px-1)+Yuy*Py*Px+Yuz*Pz*Px)*xca0
         -(Yux1*(Px*Px-1)+Yuy1*Py*Px+Yuz1*Pz*Px)*xca1
         +(Yux*Px*Py+Yuy*(Py*Py-1)+Yuz*Pz*Py)*yca0
         -(Yux1*Px*Py+Yuy1*(Py*Py-1)+Yuz1*Pz*Py)*yca1
```

```
        +(Yux*Px*Pz+Yuy*Py*Pz+Yuz*(Pz*Pz-1))*zca0
        -(Yux1*Px*Pz+Yuy1*Py*Pz+Yuz1*(Pz*Pz-1))*zca1;
A[3][1]=Px; A[3][2]=Py; A[3][3]=Pz; A[3][4]=Px*mitx+Py*mity+Pz*mitz;
DETER=A[1][1]*A[2][2]*A[3][3]+A[1][2]*A[2][3]*A[3][1]
     +A[1][3]*A[2][1]*A[3][2]-A[3][1]*A[2][2]*A[1][3]
     -A[3][2]*A[2][3]*A[1][1]-A[3][3]*A[2][1]*A[1][2];

if (DETER == (REAL4)0.0) return 0-1;

     DX=A[1][4]*A[2][2]*A[3][3]+A[1][2]*A[2][3]*A[3][4]
       +A[1][3]*A[2][4]*A[3][2]-A[3][4]*A[2][2]*A[1][3]
       -A[3][2]*A[2][3]*A[1][4]-A[3][3]*A[2][4]*A[1][2];

     DY=A[1][1]*A[2][4]*A[3][3]+A[1][4]*A[2][3]*A[3][1]
       +A[1][3]*A[2][1]*A[3][4]-A[3][1]*A[2][4]*A[1][3]
       -A[3][4]*A[2][3]*A[1][1]-A[3][3]*A[2][1]*A[1][4];

     DZ=A[1][1]*A[2][2]*A[3][4]+A[1][2]*A[2][4]*A[3][1]
       +A[1][4]*A[2][1]*A[3][2]-A[3][1]*A[2][2]*A[1][4]
       -A[3][2]*A[2][4]*A[1][1]-A[3][4]*A[2][1]*A[1][2];

*rxPs=(REAL4)(DX/DETER); *ryPs=(REAL4)(DY/DETER); *rzPs=(REAL4)(DZ/DETER);

return 0; }/* End of lohax() function */
```

“Μου έκανε το ωραιότερο δώρο που μου έχουν κάνει ποτέ:
ένα ζευγάρι παπούτσια.”

(Σε ελεύθερη απόδοση από την ταινία “Forrest Gump”)

Μέρος Β / Κεφάλαιο Ι:

Σχεδιασμός πρωτεϊνών

Εισαγωγή

Τα αποτελέσματα της ανάλυσης, που παρουσιάστηκε στα προηγούμενα κεφάλαια, δείχνουν -εμμέσως πλην σαφώς- ότι το δεμάτι μπορεί να περιγραφεί σε όρους επιμέρους (όχι όμως "ανεξαρτήτων") συνιστωσών· επίσης -συνολικά- περιέχουν τις ιδιότητες των συνιστωσών αυτών (συμπεριλαμβανομένων και των πιθανών μορφών τους) και τις μεταξύ τους σχέσεις. Όπως συμβαίνει και με τα αντικείμενα της καθημερινότητας, μόλις υπάρξει γνώση αυτού του είδους, η πρώτη και αμεσότερη εφαρμογή, που έρχεται στον δημιουργικό και γεμάτο φαντασία ανθρώπινο νου, είναι οι δυνατότητες βελτίωσης των επιμέρους συνιστωσών, με στόχο ένα πιο χρήσιμο σύνολο, ή/και ο ανασυνδυασμός τους σε νέα σύνολα.

Το παρόν κεφάλαιο αναφέρεται σε τρόπους, με τους οποίους γνώση εξειδικευμένη για ένα δομικό πρότυπο μπορεί να βοηθήσει πειράματα μεταλλαξογένεσης, είτε σημειακά είτε προκειμένου να δημιουργηθεί ένα ολόκληρο νέο πρωτεϊνικό μόριο (που με τη σειρά του θα μπορούσε να αποτελέσει την πλατφόρμα για παραπέρα πειράματα) και αναλύεται σε λεπτομέρεια μια τέτοια περίπτωση. Όμως, εξαιτίας της φύσης του κεφαλαίου, το ογκωδέστερο αλλά και σημαντικότερο τμήμα είναι -φυσικά- η συζήτηση. Η ανάλυση και ο προσχεδιασμός του μορίου, όπως περιγράφονται, έγιναν στο τέλος του 1992, με την ελπίδα ότι συνεργαζόμενη ομάδα στην Ιταλία θα αναλάμβανε το πειραματικό μέρος, κάτι που τελικά δεν έγινε.

Γενικά

Κυριολεκτικά, ο όρος *σχεδιασμός πρωτεϊνών* (protein design) αναφέρεται στο σχεδιασμό από το μηδέν μιας πολυπεπτιδικής αλυσίδας, που δεν σχετίζεται με οποιοδήποτε τρόπο με τις ήδη γνωστές, και που θα διπλώσει σε μια *επακριβώς* προκαθορισμένη δομή. Μπορεί ακόμη να καλύπτει έναν ευρείας κλίμακας επανασχεδιασμό μιας σημαντικής συνιστώσας, ώστε το μόριο να αποκτήσει μια εντελώς καινούρια ιδιότητα. Σημαίνει λοιπόν -έμμεσα- ότι γνωρίζει κανείς τους κανόνες, με βάση τους οποίους σχεδιάζονται πρωτεΐνες, και προχωρά σε εφαρμογές. *Στην πράξη*, έχει χρησιμοποιηθεί για να καλύψει διάφορες δραστηριότητες, κυρίως τέτοιες που αποσκοπούν μάλλον στην απόκτηση σχετικής γνώσης, ή -σε μια παραλλαγή- στη διαπίστωση κατά πόσο η υπάρχουσα γνώση επαρκεί, παρά στην εφαρμογή με στόχο συγκεκριμένα προϊόντα. Όταν πρόκειται για σχεδιασμό μιας ολόκληρης πρωτεΐνης, συνήθως επιλέγεται κάποιο δομικό πρότυπο που να μπορεί να υλοποιηθεί σαν μια αλληλουχία αρκετά μικρή ώστε να είναι προσπελάσιμη πειραματικά, αλλά αρκετά μεγάλη ώστε να δίνει ένα σταθερό μόριο (ιδανικά γύρω στα 70 κατάλοιπα, και πάντως κάτω από τα 100). Ο στόχος είναι είτε η μελέτη των αρχών που διέπουν το δίπλωμα σε ένα γενικό επίπεδο, είτε -πιο πρόσφατα- η δημιουργία δυνητικά χρήσιμων προϊόντων. Έτσι, θα πρέπει το δομικό πρότυπο να έχει κατανοητή και κατά προτίμηση απλή δόμηση· να μπορούν να εξαχθούν κανόνες με βάση τους οποίους θα σχεδιάζονται αλληλουχίες, αφού πρώτα απ'όλα στην πράξη θα δοκιμαστούν οι κανόνες αυτοί.

> Οι περισσότερες μικρές πρωτεΐνες δεν είναι κατάλληλες από πλευράς δομικού προτύπου, αφού η δομησή τους είναι αρκετά πολύπλοκη ώστε να αναφέρονται σαν "μη κανονικές" *(small irregural proteins)*. Το δεμάτι, αντίθετα, ικανοποιεί όλα τα παραπάνω κριτήρια: τέσσερις α-έλικες με μήκος 4-6 στροφές και σύντομες συνδέσεις δεν υπερβαίνουν τα 80 κατάλοιπα· οι στατιστικές των α-ελίκων είναι (όπως είδαμε) εξαιρετικά σαφείς (σε επόμενο κεφάλαιο θα δειχτεί και πως διαφέρουν από εκείνες συναφών προτύπων), όπως και των συνδετικών τμημάτων, που -επιπλέον- η γεωμετρία τους, καθώς και οι ακτινικές θέσεις όπου αρχίζουν και τελειώνουν κατά προτίμηση, είναι συγκεκριμένα και περισσότερο από κατανοητά. Το σημαντικότερο: δείχνει μεγάλα περιθώρια "ανοχής", σε επίπεδο αλληλουχίας, δηλαδή υπάρχει τόσο μια ποικιλία αλληλουχιών μέσα στην κάθε πρωτεϊνική οικογένεια, όσο -κυρίως!- μια ποικιλία διαφορετικών οικογενειών που το υιοθετούν, ένα ακόμη γεγονός που δείχνει πόσο γενικής φύσεως είναι οι περιορισμοί που θέτει ένα δομικό πρότυπο. Αυτά τα χαρακτηριστικά το κάνουν ιδανικό τόσο για πειράματα de novo σύνθεσης ολόκληρων πρωτεϊνών, όσο και για πειράματα συστηματικών σημειακών αλλαγών, προκειμένου να διερευνηθεί θέση προς θέση η σπουδαιότητα των επιμέρους συνιστωσών στην σταθερότητά του. Η μεγάλη ανοχή σε αλληλουχίες δεν αποκλείεται να γενικεύεται για τα λίγα δομικά πρότυπα στα οποία διπλώνουν οι πρωτεΐνες (έστω για τα πιο "κανονικά" από αυτά), και τα οποία αποτελούν εξελικτικά επελεγμένες "βιώσιμες" λύσεις (δηλαδή θερμοδυναμικά και λειτουργικά σταθερές).

Όμως, το μέγεθος μιας πολυπεπτιδικής αλυσίδας 70-100 καταλοίπων, θέτει κάποια σοβαρά προβλήματα. Μια πολυπεπτιδική αλυσίδα, με μήκος μεγαλύτερο από 30-50 κατάλοιπα, μπορεί και

διπλώνει σε -κάποια- σταθερή δομή[24]. Στις φυσικού τύπου πρωτεΐνες, αυτή η δομή έχει ρυθμιστεί εξελικτικά, όχι για τη μεγαλύτερη σταθερότητα, αλλά για την αποτελεσμα-τικότερη βιολογική λειτουργία, που όμως συμβαδίζει με σχετικά μικρή θερμοδυναμική σταθερότητα [5-10 Kcal/mol -Dill, 1990], της οποίας ο από πριν θεωρητικός χειρισμός είναι δύσκολος. Εκτός αυτού, ένα τέτοιο μέγεθος είναι δυσπρόσιτο στις μεθόδους σύνθεσης· οι τεχνικές του ανασυνδυασμένου DNA αίρουν (σταδιακά) *εν μέρει* μόνο τον περιορισμό αυτό, αφού η εισαγωγή στοιχείων στην αλυσίδα, άλλων από τους 20 συνήθεις αμινοξικούς τύπους, δεν είναι άμεση, ούτε η χημική τροποποίηση πλευρικών αλυσίδων εύκολη. Όχι τυχαία, λοιπόν, ένας μεγάλος όγκος προσπαθειών σχεδιασμού εμπλέκει (ολιγο)πεπτίδια μικρού-μεσαίου μεγέθους (μέχρι περίπου 25 κατάλοιπα), όπου συχνά μεταβάλλονται συστηματικά κάποιες υπό ανάλυση θέσεις. Ας σημειωθεί, ότι ολιγοπεπτίδια αυτών των μεγεθών, σαν μονομερή σε διάλυμα, δεν έχουν -κατά κανόνα- συγκεκριμένη δομή· όταν αλληλεπιδρούν με άλλα μόρια (πχ. υποδοχείς ή μεταξύ τους για να σχηματίσουν ολιγομερή), η δομή τους υπαγορεύεται από το περιβάλλον με το οποίο αλληλεπιδρούν. Αυτό απαλλάσσει την πειραματική διαδικασία από τις δομικές επιπτώσεις που -τυχόν- θα είχαν οι αλλαγές, αν το πεπτίδιο μπορούσε να έχει από μόνο του μια διαμόρφωση και αλληλεπιδρούσε όντας σε αυτή. Από την άλλη πλευρά, αν κάποια διαμόρφωση είναι προτιμητέα, τότε ένα πλήθος αναλόγων με περιορισμένη ελευθερία περιστροφής γύρω από συγκεκριμένους δεσμούς είναι διαθέσιμα (πχ υποκατεστημένα στον α-άνθρακα, πεπτιδικά ή μη, ή κυκλοποιημένα κ.α.) και μπορούν να ενσωματωθούν εύκολα.

Η (σχετική) ευκολία σύνθεσης και χρήσης των (μικρότερων) πεπτιδίων και η συνήθως ευχερέστερη εξαγωγή συμπερασμάτων, ειδικά καθώς τα αποτελέσματα μπορούν να αποτιμηθούν και θεωρητικά (οι ενεργειακοί υπολογισμοί είναι αρκετά ακριβείς για τέτοια συστήματα), τα καθιστούν ελκυστικά σαν συστήματα-μοντέλλα (εφ'όσον είναι δυνατή η χρήση τους), όταν επιλέγονται προσεγγίσεις με συνθετικά, τεχνητά ή/και τροποποιημένα μόρια. Επίσης, διευκολύνουν και από την άποψη ότι το μόριο που θα σχεδιαστεί θα πρέπει να είναι (*ει δυνατόν*) πρότυπο, κάτι που δύσκολα επιτυγχάνεται δουλεύοντας στα πλαίσια μιας (συγκεκριμένης) πρωτεϊνικής οικογένειας. Έχουν χρησιμοποιηθεί ευρύτατα σε μελέτες λειτουργίας (πχ πρόσδεση σε υποδοχείς, σαν βιολογικά αδρανή ανάλογα συνήθως) αλλά και γενικότερων αρχών της δομής των πρωτεϊνών.

Όμως, η έλλειψη λεπτομερών κανόνων συσχετισμού δομής-αλληλουχίας έχει περιορίσει τις δυνατότητες τέτοιων διαδικασιών: στην πράξη, συνήθως βασίζονται σε γενικής φύσεως πληροφορία, όπως προτιμήσεις των διαφόρων αμινοξικών τύπων για β-ταγή στοιχεία, ενώ η "προκαθορισμένη" δομή είναι συνήθως επίσης γενικά ορισμένη. Τα μόρια παίρνουν τη μορφή ολιγομερών, που αποτελούνται από ολιγοπεπτίδια (ώστε να διευκολύνεται η σύνθεση) με απλές αλληλουχίες (συχνά με εσωτερικές επαναλήψεις). Τα επιμέρους ολιγοπεπτίδια συνήθως σχεδιάζονται να αλληλεπιδρούν μεταξύ τους σε διαμόρφωση α-έλικας, ενώ συχνά χρησιμοποιούνται μέταλλα ή σύμπλοκά τους για τη

[24] Αυτή μπορεί να είναι είτε μια και συγκεκριμένη (αν η πρωτεΐνη είναι φυσικού τύπου ή συμπεριφέρεται έτσι), είτε ένα συνοθύλευμα από δομές (αν λείπει η δυνατότητα συμπεριφοράς φυσικού τύπου).

σταθεροποίηση του τελικού μορίου. Κενό υπάρχει και στις διαδικασίες εκ των προτέρων εκτίμησης του μοντέλου (πχ. κατά πόσο θα διπλωθεί καν), ενώ συστήματα με αντίστοιχες απλοποιήσεις (πχ α-έλικες που αποτελούνται μόνο από αλανίνη -δες Συζήτηση) συχνά χρησιμοποιούνται και για θεωρητικούς υπολογισμούς. Αν και χρησιμοποιούνται διάφορες βιοχημικές και βιοφυσικές προσεγγίσεις, στο κάθε βήμα της πειραματικής διαδικασίας, για να ελεγθεί αν συμβαδίζει με τα θεωρητικώς αναμενόμενα, η δομική ανάλυση, που είναι η μόνη που μπορεί να κρίνει τελεσίδικα την έκβαση του εγχειρήματος, συχνά οδηγεί σε εκπλήξεις, αφού -τελικά όχι απρόσμενα- συχνά το αποτέλεσμα συμφωνεί μόνο σε γενικά χαρακτηριστικά με το ζητούμενο.

> Πλήρης κάλυψη του θέματος στο παρόν πλαίσιο είναι αδύνατη και εκτός σκοπού, αφού η διαδικασία σχεδιασμού που περιγράφεται στη συνέχεια είναι διαφορετικού τύπου. Για μια εκτενέστερη ανάλυση της φιλοσοφίας πίσω από το σχεδιασμό μικρών πεπτιδίων, καθώς και μια σειρά από λεπτομερή παραδείγματα, ο αναγνώστης μπορεί να αναφερθεί σε σχετικό άρθρο του DeGrado [1988], ενώ γενικότερα για θέματα σχεδιασμού, στο άρθρο ανασκόπησης των DeGrado et. al [1999]. Εδώ αξίζει να αναφερθεί -μόνο- ότι, μια από τις πρώτες προσπάθειες σχεδιασμού ολόκληρων πρωτεϊνικών μορίων, αφορούσε ένα μικρό πεπτίδιο με μήκος 24 κατάλοιπα, το πεπτίδιο πρόσδεσης DDT [Moser et al, 1983, 1987]. Ο στόχος ήταν ένα αντι-παράλληλο β-πτυχωτό φύλλο από τέσσερις β-κλώνους, ανοικτό (εκτεθειμένο) και από τις δύο πλευρές, όπου βρίσκονται οι (υδρόφοβες) θέσεις πρόσδεσης του DDT. Προβλήματα στη διαλυτότητα δεν επέτρεψαν δομική ανάλυση, παρ'ότι βιοχημική δουλειά έδειξε ότι προσδένει το DDT 140 φορές ισχυρότερα από ότι τυχαίες αλληλουχίες με το ίδιο μήκος και την ίδια σύσταση.

Όσον αφορά τη δημιουργία δυνητικά χρήσιμων προϊόντων αυτή μπορεί να έχει τη μορφή εντοπισμένων παρεμβάσεων. Μια αλλαγή (μια ή μερικές πλευρικές αλυσίδες ή/και ενδεχόμενα ένα συνδετικό τμήμα), που θα δημιουργήσει μια <u>νέα</u> θέση πρόσδεσης για ένα μέταλλο ή ένα ενεργό κέντρο εκεί όπου δεν υπήρχε, αποτελεί επίσης διαδικασία σχεδιασμού[25] (σαν τμήμα μια διαδικασίας protein engineering).

Όμως, στην πράξη, και αυτή η κατηγορία δραστηριοτήτων αποσκοπεί κυρίως στην απόκτηση σχετικής γνώσης. Στην πλειοψηφία τους, παίρνουν τη μορφή μιας σειράς πειραμάτων "δοκιμής και σφάλματος" με απλές, αλλά συστηματικές αλλαγές, εντοπισμένες σε ένα σημείο ή μια μικρή περιοχή της αλληλουχίας ή της δομής: ένα ή λίγα αμινοξικά κατάλοιπα μεταλλάσσονται κάθε φορά (ή/και προσθαφαιρούνται κάποια τμημάτα), συνήθως με κατευθυνόμενη μεταλλαξογένεση και προσδιορίζονται οι επιπτώσεις στη σταθερότητα (συνήθως θερμιδομετρικά ή/και φασματοσκοπικά) και λειτουργικότητα (συνήθως σαν αλλαγή στη συγγένεια προς ένα υπόστρωμα, συν τις σχετικές κινητικές) του μορίου. Συνήθως στοχεύουν στον εντοπισμό και στη μελέτη καταλοίπων "κλειδιών" για τη σταθερότητα ή τη λειτουργία της πρωτεΐνης, ή στην απόκτηση νέων ιδιοτήτων από το μόριο. Συνήθως αποτελούν μέρος γενικότερων βιοχημικών και βιοφυσικών προσεγγίσεων, των οποίων είτε καθοδηγούν το πειραματικό μέρος στα πλέον υποσχόμενα μονοπάτια, είτε βοηθούν στο να εξηγηθούν (άλλα)

[25] Η αλλαγή όμως ενός καταλοίπου στο ενεργό κέντρο, ώστε να αλλάξει η ειδικότητα του ενζύμου, όχι (αν και είναι πάλι τμήμα διαδικασίας protein engineering).

πολύπλοκα πρωτογενή πειραματικά δεδομένα. Πράγματι, οι παραπάνω διαδικασίες (συνήθως και σε συνδυασμό με ακόλουθη ανάλυση με φασματοσκοπία ή κρυσταλλογραφία), έχουν χρησιμοποιηθεί εκτενώς στο παρελθόν για την κατανόηση της λειτουργίας πολλών πρωτεϊνών, ενώ στο μέλλον αναμένεται να βοηθήσουν στο σχεδιασμό νέων, με χρήσιμες ιδιότητες. Και φυσικά, αν και εντάσσονται σαφώς στα πλαίσια της πρωτεϊνικής μηχανικής, ως διαδικασίες σχεδιασμού μπορούν να θεωρηθούν μόνο με τον διευρυμένο ορισμό που δόθηκε στην αρχή.

Προκειμένου για τη μελέτη των ιδιοτήτων σε μια ή πολλαπλές θέσεις-κλειδιά με συστηματικές αλλαγές, καθώς οι δομικές ιδιότητες κάθε θέσης είναι δύσκολο να προβλεφθούν, και σε συνδυασμό με την ελλειπή γνώση των κανόνων του διπλώματος, μικρές, φαινομενικά "ανώδυνες" αλλαγές μπορούν να οδηγήσουν σε ένα μόριο που δεν διπλώνει. Έτσι, η πρόβλεψη των πιθανών αποτελεσμάτων πολλαπλών ή ευρείας κλίμακας τροποποιήσεων είναι δυσχερής, ενώ συχνά υπάρχει σύγχυση ακόμα και κατά την ερμηνεία των πειραματικών εξαγομένων: *ακόμη και για μια απλή μεταλλαγή, μιας υποτιθέμενης βιολογικά ενεργής θέσης, η έλλειψη βιολογικής δραστικότητας, που συνήθως χρησιμοποιείται σαν δείκτης, μπορεί να οφείλεται τόσο σε λειτουργική όσο όμως και σε δομική αναγκαιότητά της. Όμως, ακόμη κι αν ήταν εύκολη η αποτίμησή τους, μικρές, εντοπισμένες (στο χώρο) αλλαγές, ξεκινώντας από μια φυσική πρωτεΐνη και παραμένοντας στο πλαίσιο, που έχει οριστεί εξελικτικά, για την συγκεκριμένη κάθε φορά πρωτεϊνική οικογένεια, δύσκολα μπορούν να οδηγήσουν σε μόρια μοντέλλα, σχεδιασμένα να μπορούν να απαντήσουν σε συγκεκριμένα ερωτήματα, γενικότερης φύσεως, ανεξάρτητα από τις ιδιαιτερότητες κάθε οικογένειας.*

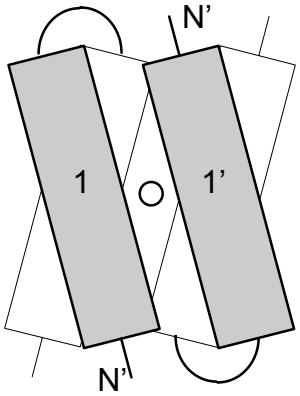
Πίνακες κατανομών, σαν εκείνους των προηγουμένων κεφαλαίων, προσφέρουν ακριβώς εξειδικευμένη πληροφορία, σχετικά με την συμβατότητα συγκεκριμένων αμινοξικών τύπων με συγκεκριμένες θέσεις ενός δομικού προτύπου, και μπορούν να διευκολύνουν όλο το φάσμα των παραπάνω δραστηριοτήτων. Πιο συγκεκριμένα, με την εφαρμογή πιο εξειδικευμένων περιορισμών, μπορούν να τεθούν πιο συγκεκριμένοι στόχοι από το σχέδιο, ενώ αυξάνουν οι πιθανότητες το αποτέλεσμα να συμφωνεί, όχι μόνο σε γενικά χαρακτηριστικά με το ζητούμενο, αλλά σε μια ικανοποιητική (ανάλογα με το στόχο) λεπτομέρεια. Η χρησιμότητα τέτοιων πινάκων κατανομών, τόσο για σημειακές παρεμβάσεις, όσο για σχεδιασμό ολόκληρων μορίων, ίσως είναι ήδη αντιληπτή, αφού προφανώς μπορούν να κατευθύνουν τέτοιες διαδικασίες στα πλέον υποσχόμενα μονοπάτια και να διευκολύνουν την ερμηνεία των εξαγομένων τους. Καθώς δε, τα (συνήθη) δομικά πρότυπα είναι λίγα, η ανάλυση αρκεί να γίνει μια φορά και να αποθηκευτούν τα αποτελέσματα σε κάποιο είδος τράπεζας πληροφοριών, τουλάχιστο για τα συχνότερα απαντώμενα από αυτά.

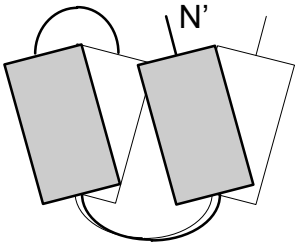
Προκειμένου όμως να συζητηθούν πιο συγκεκριμένα (με βάση ένα εκτενές παράδειγμα) τα πλεονεκτήματα, που προκύπτουν από τη χρήση πληροφορίας εξειδικευμένης για κάθε δομικό πρότυπο, περιγράφεται στη συνέχεια μια διαδικασία σχεδιασμού, που εκμεταλλεύεται τα δεδομένα των προηγουμένων κεφαλαίων, και άπτεται πολλών από τα θέματα που συζητήθηκαν μέχρι εδώ. Κάτι δε, που πρέπει να προσεχτεί ιδιαίτερα, είναι το γεγονός ότι -*αν και δεν σχεδιάζεται από το μηδέν καμία συνιστώσα του τελικού προϊόντος- το μόριο που προκύπτει αποτελεί μοντέλλο,* όπως διευκρινίζεται σε άλλο σημείο, πιο πρόσφορο για μια τέτοια εξήγηση. Η συζήτηση κλείνει με την τοποθέτηση της συγκεκριμένης διαδικασίας στο γενικότερο πλαίσιο, όπου επιπλέον γίνονται σαφή τα όρια των δυνατοτήτων, για πίνακες του είδους, καθώς και τα επιπλέον στοιχεία που απαιτούνται ώστε η διαδικασία να είναι πλήρης.

Διαδικασία

Ανάλυση του ζητήματος: Η πρωτεΐνη ROP σχηματίζεται σαν ένα ομοδιμερές, στο οποίο τα επιμέρους μονομερή Α και Β σχετίζονται με ένα ακριβή (κρυσταλλογραφικό) δυαδικό άξονα συμμετρίας (Εικόνα 1α) [Banner et al, 1987]. Οι αλληλεπιδράσεις μεταξύ των α-ελίκων, παρουσιάζουν μια χαρακτηριστική στρωμάτωση σε οκτώ"φέτες" κάθετες στην κατά μήκος των αξόνων των α-ελίκων διεύθυνση του δηματιού (Εικόνα 7/Γεν. Εισαγωγή). Εξαιτίας της συμμετρίας, το αμινοτελικό μισό της πρώτης α-έλικας του μονομερούς Α αντικρύζει το καρβοξυτελικό μισό της ίδιας α-έλικας από το μονομερές Β. Επίσης, το καρβοξυτελικό μισό της δεύτερης α-έλικας του μονομερούς Α αντικρύζει το αμινοτελικό μισό της ίδιας α-έλικας από το μονομερές Β. Έτσι, θεωρητικά, αν κανείς μπορούσε να τροποποιήσει τα μέσα των α-ελίκων του ενός μονομερούς, ώστε η κάθεμια από τις δύο α-έλικες να διπλωθεί στη μέση σχηματίζοντας μια φουρκέττα, το αποτέλεσμα θα ήταν ένα δεμάτι αποτελούμενο από τέσσερις "μισές" α-έλικες, όλες προερχόμενες από ένα μονομερές (Εικόνα 1β). Η τροποποίηση θα μπορούσε να περιλαμβάνει -εκτός από σημειακές μεταλλαγές- και εισδοχή λίγων καταλοίπων ή/και ενός συνδετικού τμήματος (όχι μεγάλου μήκους) από κάποιο άλλο δεμάτι, χωρίς όμως -σε κάθε περίπτωση- να διακόπτεται η συνέχεια του προτύπου ή να αλλοιώνεται ο χαρακτήρας του.

Εικ. 1α*. Σχηματική αναπαράσταση της ομοδιμερούς πρωτεΐνης ROP φυσικού τύπου.Ο άξονας συμμετρίας, που μας ενδιαφέρει στο πείραμα, είναι κάθετος στη σελίδα, και διέρχεται μεταξύ των α-ελίκων 1 και 1' (στο κέντρο του σχήματος). Περιστροφή του αμινοτελικού μισού της α-έλικας 1, γύρω από αυτόν τον άξονα, δίνει το αμινοτελικό μισό της α-έλικας 1'. Αντίστοιχα ισχύουν για τις α-έλικες 2 και 2', που φαίνονται με αχνές γραμμές στο "πίσω" μέρος.*





Εικ. 1β*. Σχηματική αναπαράσταση του επιθυμητού τελικού μορίου. Το αμινοτελικό άκρο που είναι σημειωμένο με Ν' είναι εκείνο της α-έλικας 1 (και όχι της 1') της Εικόνας 1α. Αντίστοιχα ισχύουν για το καρβοξυτελικό άκρο της α-έλικας 2 στο "πίσω" μέρος. Και οι τέσσερις μισού μήκους α-έλικες προέρχονται από ένα μονομερές της πρωτεΐνης ROP φυσικού τύπου. Εδώ δεν υπάρχει πλέον άξονας συμμετρίας.*

Αρχικά η διαδικασία αυτή καταστρώθηκε για λόγους επίδειξης, γύρω από το υποθετικό τότε ζητούμενο -σε συνδυασμό και με άλλα δυνητικά πειράματα, που παρουσιάζονται παρακάτω συνοπτικά- της επάρκειας της πληροφορίας των επτάδων των θέσεων για σχηματισμό δηματιού, δεδομένου ενός

υδρόφοβου πυρήνα μισού μεγέθους (σημ.: η φυσική πρωτεΐνη ROP είναι ιδιαίτερα σταθερή), και της μεταξύ τους ανεξαρτησίας.

Σχεδιασμός του μορίου: Στη φυσική πρωτεΐνη ROP, η πρώτη α-έλικα του μονομερούς περιλαμβάνει αμινοξικά κατάλοιπα όχι ευνοϊκά για α-έλικες, όπως φαίνεται από το γεγονός ότι όλες οι κλασσικές μέθοδοι πρόβλεψης (όπως υλοποιούνται στο πρόγραμμα PREDICT) [CCP4, 1979 και εκεί αναφορές] ανιχνεύουν ασυνέχειες στην α-έλικα αυτή (Εικόνα 2). Μάλιστα, υπάρχει πραγματικά μια έντονη κάμψη της έλικας αυτής στο σημείο αυτό, οι ιδιαιτερότητες του οποίου έχουν σχολιασθεί στο παρελθόν [Banner et al, 1987]. Η δεύτερη α-έλικα, από την άλλη, δεν δείχνει τέτοιες "αδυναμίες". Έτσι, ίσως είναι αρκετό να γίνουν κάποιες αλλαγές στο μέσο της δεύτερης α-έλικας, με την πρώτη να ακολουθεί, απλά υποχωρώντας στο "αδύνατο σημείο". Τα συνδετικά τμήματα που απαιτούνται, για να συνδεθούν οι τέσσερις μισές α-έλικες (στα τέσσερα ελεύθερα άκρα που προκύπτουν, ανά δύο) φαίνονται στο κάτω μέρος της Εικόνας 1β, είναι δε, σύμφωνα με τους ορισμούς του Κεφ. Α.ΙΙ, αριστερόστροφα. Καθώς είναι καλό να διατηρηθούν οι αλληλεπιδράσεις της "φέτας" υπ' αριθμό 4 (Εικόνα 7/Γεν. Εισαγωγή), θα πρέπει -στο τελικό μόριο- τα τμήματα αυτά να ξεκινούν μετά από κάποια θέση d, και να τελειώνουν πριν από κάποια θέση a, ενώ τα άκρα τους θα πρέπει να είναι περίπου στο ίδιο "ύψος" στη διεύθυνση κατά μήκος των α-ελίκων του δεματιού.

```
                    ***** JOINT PREDICTION RESULTS *****
                                ROP PROTEIN

  NOTE:
  =====
  SMALL LETTERS (h,b,t) INDICATE PREDICTION BY LESS THAN 5 METHODS
  LARGE LETTERS (H,B,T) INDICATE PREDICTION BY 5 METHODS AND ABOVE

         1         2         3         4         5         6
    5    0    5    0    5    0    5    0    5    0    5    0
MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARFGDDGENL
  hhHHHHHHHHHHH     HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHh          h
                   bB                                  b
                tt                                     tTTTTt
```

Εικ. 2*. Απόσπασμα (περίληψη) των αποτελεσμάτων του προγράμματος PREDICT για την αλληλουχία της πρωτεΐνης ROP. Το πρόγραμμα χρησιμοποιεί συνολικά οκτώ μεθόδους πρόβλεψης και δίνει σαν περίληψη την εκτίμηση των περισσοτέρων μεθόδων. Είναι χαρακτηριστική η έλλειψη προβλεπομένων στοιχείων β-πτυχωτών φύλλων. Η περιοχή γύρω από την αλανίνη στη θέση 31, αν και προβλέπεται ελικοειδής, αποτελεί το συνδετικό τμήμα μεταξύ των ελίκων 1 και 2.*

Με αυτές τις παραμέτρους κατά νου, αρχικά η αναζήτηση στράφηκε στα ήδη γνωστά δεμάτια, για κατάλληλα συνδετικά τμήματα, και πρώτα για τη δεύτερη α-έλικα, αφού αυτή είναι ο κύριος στόχος του πειράματος. Με εκτεταμένη χρήση του Πίνακα Α.ΙΙ.2 προκύπτουν τρεις κατάλληλοι υποψήφιοι: το συνδετικό τμήμα μεταξύ των α-ελίκων 1 και 2 της πρωτεΐνης του καλύμματος (coat protein) του ιού της μωσαϊκής του καπνού (2TMV(1->2)), και εκείνα των α-φουρκετών 1-2 και 2-3 της αιμερυθρίνης

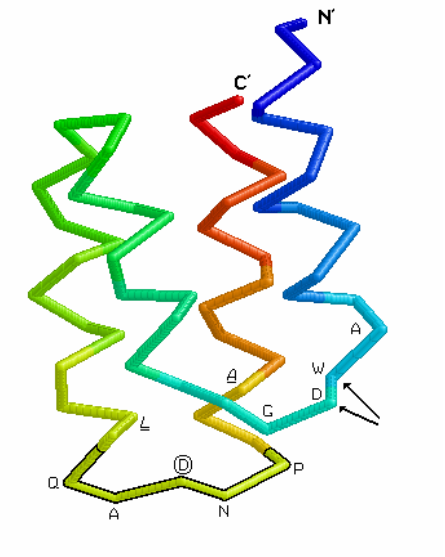
(2HMZ(1->2) και 2HMZ(2->3)). Αν επιλεγεί το τμήμα 2TMV(1->2), πρέπει να γίνει μια εισδοχή δέκα αμινοξικών καταλοίπων. Το τμήμα 2HMZ(2->3) είναι δεξιόστροφο, αν και κατάλληλη μετάλλαξη κατά μήκος του θα μπορούσε να διορθώσει την εναλλαγή υδρόφοβων-πολικών καταλοίπων. Όμως το τμήμα 2HMZ(1->2) (αριστερά στην Εικόνα Α.Ι.4), προσέφερε μια ακόμη καλύτερη λύση.

Πιο συγκεκριμένα, το κατάλοιπο 35 της αιμερυθρίνης (κωδικός PDB: 2HMZ) είναι μια λευκίνη (35L) σε θέση d, ακολουθούμενη από την αλληλουχία 36-SQADNADHL-44, όπου η τελευταία λευκίνη (44L) είναι σε θέση a, (για τα υπογραμμισμένα κατάλοιπα δες παρακάτω). Η πρωτεΐνη ROP έχει στη θέση 41 (μετά την οποία κάπου πρέπει να τσακίσει η έλικα) επίσης ένα κατάλοιπο λευκίνης (41L) σε θέση d, ακολουθούμενη από την αλληλουχία 42-HDHA-45, με την τελευταία αλανίνη (45A) επίσης σε θέση a. Συστοιχίζοντας στο χώρο την 35L(2HMZ) με τη 41L(ROP), την 36S(2HMZ) με τη 42H(ROP), τη 42D(2HMZ) με τη 43D'(ROP -το άλλο αρχικό μονομερές), τη 43H(2HMZ) με τη 44H'(ROP), και τη 44L(2HMZ) με τη 45A'(ROP), προκύπτει ότι η εισδοχή της υπογραμμισμένης αλληλουχίας (QADNA) αποτελεί μια πιθανή λύση στο σκέλος αυτό του προβλήματος, αφού γίνεται ανάμεσα σε μια θέση d και μια θέση a (στο τελικό μόριο), που προέρχεται επίσης από θέση a (από το αρχικό μόριο), διατηρώντας έτσι το σωστό πλαίσιο ανάγνωσης των ακτινικών θέσεων, ενώ συντηρούνται και τρία από τα πέντε συστοιχιζόμενα κατάλοιπα. Ακόλουθη αλλαγή της 45A(ROP) σε λευκίνη, κατά τη δημιουργία του μοντέλλου, αύξησε τον αριθμό σε 4/5 . Η εισδοχή πέντε καταλοίπων βγάζει την επανάληψη επτάδας μεταξύ των 41L(ROP) και 45A(ROP) εκτός φάσης κατά δύο θέσεις (200^{O} γύρω από τον άξονα της έλικας), κάτι που όπως έχει ήδη σχολιαστεί κάνει αδύνατο το σχηματισμό μιας συνεχόμενης υδρόφοβης ταινίας, και συνεπώς μιας *ενιαία* αμφιπαθικής α-έλικας. Για να δυσκολέψει περισσότερο ο σχηματισμός μιας συνεχόμενης α-έλικας, το τελευταίο κατάλοιπο αλανίνης στην εισδοχή QADNA, καθώς αναμένεται στο τελικό μόριο να καταλάβει μια θέση $N_{άκρο}$+1, μπορεί να αλλαχτεί σε προλίνη, που έχει πολύ μεγάλη συμβατότητα με τη θέση αυτή (Πίνακας Α.ΙΙ.4). Έτσι, η προτεινόμενη εισδοχή για το μέσο της δεύτερης α-έλικας γίνεται QADNP. Εξετάζοντας οπτικά -σε οθόνη γραφικών και με τις πρωτεΐνες ROP και αιμερυθρίνη σε αλληλυπέρθεση- τον υποκείμενο υδρόφοβο πυρήνα, αντίκρυ στον οποίο θα πακεταριστεί το συνδετικό τμήμα, προκύπτουν ομοιότητες με τον αντίστοιχο στην αιμερυθρίνη, αντίκρυ από τον οποίο βρίσκεται το συνδετικό τμήμα που "αντιγράφτηκε".

Απλές συμπτώσεις, ή καλός καθορισμός του δεματιού;

Ακολουθούν οι τροποποιήσεις που αφορούν το μέσο της πρώτης α-έλικας. Εδώ, οποιαδήποτε αλλαγή θα πρέπει να γίνει μεταξύ της ισολευκίνης στη θέση 15 (15I), που θα είναι η τελευταία θέση d πριν το συνδετικό τμήμα, και της θρεονίνης στη θέση 19 (19T), που θα είναι και η πρώτη θέση a αμέσως μετά (δες και Εικόνα 7/Γεν. Εισαγωγή). Μια προφανής πιθανότητα είναι το συνδετικό τμήμα μεταξύ των ελίκων 3 και 4 της αιμερυθρίνης (HMZ(3->4)) (δεξιά στην Εικόνα Α.Ι.4), το οποίο στο φυσικό του περιβάλλον βρίσκεται απέναντι από το συνδετικό τμήμα που χρησιμοποιήθηκε παραπάνω.

Έτσι, το περιβάλλον, τόσο για το ίδιο όσο και για το άλλο συνδετικό τμήμα διατηρείται ακόμη περισσότερο.



Εικ. 3. *Σχηματική αναπαράσταση του αναμενόμενου τελικού μορίου, στον ίδιο περίπου προσανατολισμό με την Εικόνα 1β (το άνω μέρος έχει "γείρει" λίγο προς τα πίσω, ώστε να ξεχωρίζουν καλύτερα τα υπό σχεδιασμό συνδετικά τμήματα στο κάτω μέρος). Οι τέσσερις α-έλικες προέρχονται από ένα μονομερές. Έτσι, το αμινο- και το καρβοξυ-τελικό άκρο που δείχνονται αντιστοιχούν σε εκείνα της αλυσίδας 1 (και όχι της 1') της Εικόνας 1α κατόπιν τσακίσματος των ελίκων στα δύο και αναδίπλωσης. Αυτό επιτυγχάνεται με μια εισδοχή στην καρβοξυτελική α-έλικα και μερικές αλλαγές στην αμινοτελική, ώστε να υπάρξει μίμηση των συνδετικών τμημάτων 2HMZ(1-2) και 2HMZ(3-4). Η υπογραμμισμένη λευκίνη είναι η 41L(ROP), σε θέση d (αντιστοιχεί στην 35L(2HMZ)), και η υπογραμμισμένη αλανίνη η 45A(ROP), σε θέση a (αντιστοιχεί στην 44L(2HMZ)), ανάμεσα -κάπου- στις οποίες πρέπει να γίνει η εισδοχή των πέντε καταλοίπων, όπως αναφέρεται στο κείμενο. Η εισδοχή τελικά γίνεται μεταξύ 42H(ROP) και 43D(ROP), και δείχνεται με έντονο περίγραμμα στο κάτω μέρος, ενώ η 45A(ROP) τροποποιείται επίσης σε λευκίνη. Στο άλλο συνδετικό τμήμα, δεν υπάρχει εισδοχή ή απαλοιφή, αλλά έχουν γίνει οι εξής μεταλλάξεις: 14F(ROP)->A, 15I(ROP)->W, 16R(ROP)->D και 17S(ROP)->G. Η τρυπτοφάνη (W) και το ασπαρτικό οξύ (D), που δείχνονται με βέλη, είναι εκείνα της εικόνας Α.ΙΙ.3. Η αλανίνη που προηγείται,προέρχεται από την 14F(ROP)· αν παρέμενε φαινυλ-αλανίνη, η πλευρική της αλυσίδα θα βρίσκοταν εκεί που τώρα διέρχεται η κύρια αλυσίδα W->D->G. Η γλυκίνη, που έπεται, είναι στη θέση εκείνη που αναφέρεται στο κείμενο ότι έχει θετική γωνία φ. Τέλος, το ασπαρτικό οξύ στο κυκλάκι, είναι εκείνο της Εικόνας Α.Ι.4, και αναφέρεται στο κείμενο ότι σχηματίζει δεσμό υδρογόνου με την τρυπτοφάνη (που δείχνεται με βέλος).*

Επιπλέον έχει το "σωστό" μήκος, όχι τόσο με την έννοια της απόστασης ανάμεσα στις δύο έλικες (υπενθύμιση: οι α-έλικες της αιμερυθρίνης απέχουν μεταξύ τους μεγαλύτερες αποστάσεις), αλλά πιο πολύ με την έννοια ότι δεν χρειάζονται εισδοχές/ απαλοιφές στην αλληλουχία της πρωτεΐνης ROP, αλλά μόνο σημειακές μεταλλαγές ώστε τμήμα της πρωτεΐνης ROP να αλλάξει, μιμούμενο το τμήμα HMZ(3->4).

Ένα πρόβλημα ανακύπτει από την τρυπτοφάνη της θέσης 87 (87W) της αιμερυθρίνης, που καλείται να αντικαταστήσει την ισολευκίνη 15 της πρωτεΐνης ROP (15I(ROP)). Στο παλιό της περιβάλλον, όπως φαίνεται στην Εικόνα Α.Ι.4, η 87W έχει αρκετό χώρο για να πακεταριστεί κάτω από το συνδετικό τμήμα, ο οποίος μπορεί να μην υπάρχει στο νέο. Αντικατάσταση με φαινυλ-αλανίνη ίσως να μην είναι επιτρεπτή, αφού όπως φαίνεται στην ίδια εικόνα, το ασπαρτικό από το γειτονικό (στο χώρο) συνδετικό τμήμα QADNA δημιουργεί έναν δεσμό υδρογόνου με την πλευρική αλυσίδα της 87W, που δεν μπορεί να γίνει

με τη φαινυλαλανίνη. Σ'αυτή την περίπτωση το ασπαρτικό αυτό ίσως στραφεί στο διαλύτη για αλληλεπίδραση, αποκαλύπτοντας τμήματα του υδρόφοβου πυρήνα, κάτι που μπορεί να αποδειχτεί πολύ αποσταθεροποιητικό. Επιπλέον, ενώ η πλευρική αλυσίδα της 87W εκτελεί χρέη θέσης d (από πλευράς πακεταρίσματος) η κύρια αλυσίδα της δεν είναι σε ελικοειδή διαμόρφωση, και περνά από μια περιοχή στην οποία, στη φυσικού τύπου ROP, βρίσκεται η πλευρική αλυσίδα της φαινυλαλανίνης 14 (14F). Η 14F, στη φυσικού τύπου ROP, πακετάρεται με την 14F του άλλου μονομερούς, που στο μόριο που σχεδιάζουμε δεν υπάρχει πια. Έτσι η 14F μπορεί να αλλαχτεί σε αλανίνη. Αν και πρόκειται για θέση c, και υπάρχουν άλλοι αμινοξικοί τύποι περισσότερο συμβατοί (Πίνακες Α.Ι.3-4), εμπλέκουν φορτία (πχ γλουταμικό) και άλλα ακραία χαρακτη-ριστικά, ενώ -στα πρώτα αυτά στάδια του σχεδιασμού- είναι καλό να είναι κανείς συντηρητικός: η αλανίνη δεν έχει ακραίες ιδιότητες και είναι εξίσου συμβατή (και μάλιστα περισσότερο από το μέσο όρο) με όλες τις ακτινικές θέσεις. Ένα δεύτερο πρόβλημα με αυτό το συνδετικό τμήμα αφορά τη θέση 89 της αιμερυθρίνης, η οποία έχει θετική δίεδρη γωνία φ. Αυτό σημαίνει ότι στο υπό σχεδιασμό μόριο η θέση αυτή θα πρέπει να είναι γλυκίνη ή άλλο κατάλοιπο συμβατό με θετική φωνία φ. Πράγματι, στις αιμερυθρίνες η θέση αυτή είναι μια συντηρημένη γλυκίνη.

Έτσι, οι τροποποιήσεις που προτείνονται αρχικά είναι 14F(ROP)->A, 15I(ROP)->W, 16R(ROP)->D και 17S(ROP)->G, ώστε να μιμηθούμε το συνδετικό τμήμα HMZ(3->4).

Αν και αρχικά η διαδικασία αυτή καταστρώθηκε για λόγους επίδειξης, όπως εξελίχθηκε, αν αποδεικνύοταν επιτυχής στο πειραματικό επίπεδο, τότε θα έδινε ισχυρές ενδείξεις κατά πόσο συνδυασμός ελίκων και συνδετικών κομματιών από διαφορετικά μόρια θα μπορούσε να δώσει ένα πρωτεϊνικό μόριο που διπλώνει, δίνοντας μια ακόμη ισχυρή ένδειξη για την ανεξαρτησία (ή το καλό ταίριασμα) των απαιτήσεων των διαφορετικών τμημάτων μιας πρωτεΐνης, και ανοίγοντας τον δρόμο για τη δημιουργία πρωτεϊνών από "ανταλλακτικά".

Τον βασικό σχεδιασμό του νέου μορίου, ακολούθησε η δημιουργία μοντέλλου με τη βοήθεια ηλεκτρονικού υπολογιστή (Εικόνα 3), για να εντοπιστούν άλλα πιθανά προβλήματα. Διατηρώντας από τη διμερή πρωτεΐνη ROP όσα τμήματα θα παρέμεναν αναλλοίωτα, "χτίστηκαν" στα ενδιάμεσα κενά τα συνδετικά τμήματα που προέρχονται από την αιμερυθρίνη.

Η δημιουργία του μοντέλλου, καθώς και οι ακόλουθες ελαχιστοποιήσεις ενέργειας, και προσομοιώσεις μοριακής δυναμικής, έγιναν με το εμπορικό πρόγραμμα SYBYL (διάθεση: Tripos Associates, Munich). Αλληλυπερθέσεις, όπου χρειάστηκαν, έγιναν με τη βοήθεια του προγράμματος "O" [Jones et al, 1991], ενώ οι υπολογισμοί προσπελασιμότητας από τον διαλύτη έγιναν με τη βοήθεια του προγράμματος DSSP [Kabsh και Sander, 1983].

Δύο μικρά προβλήματα ακόμη αναφάνηκαν: με τις αλλαγές στην πορεία της κύριας αλυσίδας, μια μεγάλη τρύπα δημιουργήθηκε εσωτερικά από το συνδετικό τμήμα. Αυτό επιδιορθώθηκε αλλάζοντας το κατάλοιπο στη θέση 45 (που είναι μια θέση a) της φυσικής πρωτεΐνης ROP από αλανίνη σε λευκίνη, αφού και οι περισσότερες αιμερυθρίνες έχουν στην ίδια θέση στο χώρο μια λευκίνη, κάνοντας έτσι τον υποκείμενο υδρόφοβο πυρήνα ακόμη πιο συμβατό για τα συνδετικά τμήματα στο υπό σχεδιασμό μόριο· παράλληλα η λευκίνη είναι το πιο συμβατό κατάλοιπο για τη θέση a. Το άλλο πρόβλημα που ίσως προκύψει είναι ότι στις αιμερυθρίνες οι α-έλικες που συνδέονται με το τμήμα QADNA είναι ελαφρά περιεστραμμένες γύρω από τους άξονές τους, σε σχέση με εκείνες της πρωτεΐνης ROP, αλλά μικρές τοπικές κινήσεις σε μια πρωτεϊνική δομή μπορούν να εξομαλύνουν τέτοιες διαφορές.

Το μοντέλλο που προέκυψε υποβλήθηκε σε ελαχιστοποίηση ενέργειας και οπτική επανεξέταση για τυχόν σημαντικές μετακινήσεις τμημάτων του μορίου, με υπέρθεση του μοντέλλου τόσο επί της πρωτεΐνης ROP, όσο και επί της αιμερυθρίνης, χωρίς τελικά να παρατηρηθεί κάτι το ιδιαίτερο. Επίσης ελέγχθηκε για τυχόντα κενά στο πακετάρισμα, ή απώλεια κάποιων υδρογονοδεσμών, χωρίς πάλι να παρατηρηθεί κάτι.

Τον σχεδιασμό του μορίου ακολούθησαν διάφορες εκτιμήσεις, με πρώτη την πρόβλεψη της β'-ταγούς δομής της νέας αλληλουχίας (Εικόνα 4). Συνοπτικά παρατηρούνται:
(α) έλλειψη στοιχείων β-πτυχωτών φύλλων, (β) το μέσο της πρώτης α-έλικας εξακολουθεί να μην προβλέπεται ελικοειδές μετά τις προτεινόμενες αλλαγές (γ) η περιοχή γύρω από την αλανίνη στη θέση 31 εξακολουθεί να προβλέπεται ελικοειδής αν και αποτελεί συνδετικό τμήμα (δ) η εισδοχή στο μέσο της δεύτερης α-έλικας προβλέπεται να διακόπτει την συνέχειά της.

Ελέγχθηκε επίσης η προσπελασιμότητα των διαφόρων καταλοίπων από το διαλύτη, για να εντοπιστούν περιπτώσεις όπου -τυχόν- υδρόφοβα κατάλοιπα, απροσπέλαστα στα δύο μητρικά μόρια, ήσαν τώρα προσπελάσιμα, όμως τα αποτέλεσματα ήταν συγκρίσιμα με ό,τι ισχύει σε ROP και αιμερυθρίνη, ανάλογα με την προέλευσή του κάθε καταλοίπου.

```
                        ***JOINT PREDICTION RESULTS***
                                MODEL 1.
NOTE
====
SMALL LETTERS (h,b,t) INDICATE PREDICTION BY LESS THAN 5 METHODS
LAGRE LETTERS (H,B,T) INDICATE PREDICTION BY 5 METHODS AND ABOVE

          1         2         3         4         5         6
    5     0    5    0    5    0    5    0    5    0    5    0
MTKQEKTALNMARAWDGATLTLLEKLNELDADEQADICESLHQADNPDHLDELYRSCLARFG
    HHHHHHHhhHHH   HHHHHHHHHHHHHHHHHHHHHHHH     hHHHHHhHHhh

               t                              tttTttt           t
```

Εικ. 4*. Απόσπασμα (περίληψη) των αποτελεσμάτων του προγράμματος PREDICT για την αλληλουχία της πρωτεΐνης ROP μισού μήκους. Και εδώ παντελής έλλειψη στοιχείων β-πτυχωτών φύλλων (των οποίων η πρόβλεψη θα εμφανίζοταν μεταξύ των προβλέψεων για α-έλικα και β-στροφές), ενώ η εισδοχή (υπογραμμισμένη) προβλέπεται σαν στροφή. Σύγκρινε με Εικόνα 2, όπου οι λεπτομέρειες για το πρόγραμμα και για την περιοχή γύρω από την αλανίνη στη θέση 31.*

Τέλος, τα τρία μόρια (ROP, αιμερυθρίνη και το μοντέλλο) υποβλήθηκαν σε μια ποιοτική πιο πολύ δοκιμασία, σε όρους προσομοίωσης μοριακής δυναμικής. Ανάμεσα στις πιθανές χρήσεις, οι προσομοιώσεις αυτές έχουν χρησιμοποιηθεί και για να αποδιατάξουν (σταδιακά) πρωτεϊνικά μόρια· ξεκινώντας κανείς από πολύ χαμηλή θερμοκρασία (πχ 50 ^{O}K), με έναν αριθμό κύκλων ικανό[26] ώστε να επέλθει εξισορρόπηση *(equilibration)*, ανυψώνει σταδιακά (πχ σε βήματα 50 ^{O}K) την προσωμοιούμενη θερμοκρασία[27], μέχρι την αποδιάταξη του μοντέλλου, ελπίζοντας ότι τα πρώτα στάδια της αποδιάταξης, έχουν κάτι κοινό με τα τελευταία του διπλώματος [Finkelstein, 1997]. Στην περίπτωση αυτή όμως δεν ενδιέφερε μια λεπτομερής ανάλυση του είδους, ούτε κάποιος υπολογισμός του ενεργειακού ισοζυγίου σταθεροποίησης της διπλωμένης δομής, αλλά ένας απλός έλεγχος μήπως το μοντέλλο καταρρέει (αποδιατάσσεται), σε κάποια σημεία, ταχύτερα από ότι τα μόρια στα οποία βασίστηκε η δημιουργία του. Η μόνη διαφορά που παρατηρήθηκε, φτάνοντας στους 270-300 ^{O}K, αφορά την ιστιδίνη που προέρχεται από τη θέση 42 της πρωτεΐνης ROP, που τώρα λόγω έλλειψης περιορισμών από το πακετάρι-σμα είναι πιο ελεύθερη να περιστρέφεται. Μάλιστα, καθώς η θέση στην οποία βρίσκεται στο υπό σχεδιασμό μόριο είναι μια θέση C-3, όπου η ιστιδίνη δεν ευνοείται, ενώ, όταν η θέση αυτή είναι ταυτόχρονα και θέση e, ευνοείται η σερίνη, ίσως να πρέπει να αλλαχτεί και η ιστιδίνη αυτή σε σερίνη. (Οι αιμερυθρίνες στη θέση αυτή έχουν σερίνη!) Κατά τα λοιπά, η αποδιάταξη έρχεται στην ίδια θερμοκρασία και ξεκινά από τις άκρες των ελίκων και τα συνδετικά τμήματα.

[26] Τα εγχειρίδια χρήσης προγραμμάτων του είδους έρχονται συνήθως με πλήρη πρωτόκολλα στο θέμα αυτό.
[27] Υπάρχουν κι άλλοι "βίαοι" τρόποι επιτάχυνσης. Δες Williams M.A., Thornton J.M., και Goodfellow J.M. (1997) *Protein Engineering* **10**, 895-903 για χρήση "διευκολυνόμενης ενυδάτωσης του υδρόφοβου πυρήνα".

Συμπεράσματα-Συζήτηση

Το αποτέλεσμα της παραπάνω διαδικασίας σχεδιασμού, ισοδυναμεί με διαίρεση της πρωτεΐνης ROP σε δύο μισά. Στη συγκεκριμένη περίπτωση, η εσωτερική συμμετρία βοήθησε, ώστε αυτό να μπορεί να υλοποιηθεί με μια μικρή εισδοχή στη φυσικού τύπου αλληλουχία και κάποιες σημειακές παρεμβάσεις· όμως το αποτέλεσμα είναι μια δραστική αλλαγή στη δομή. Τελικά, ελέγχεται *συνολικά* η επάρκεια της πληροφορίας των πινάκων, για ένα σταθερό μόριο (δεδομένου ενός υδρόφοβου πυρήνα μισού μεγέθους), και όχι μόνο των επτάδων (όπως αρχικά ήταν ο στόχος), αφού -αρκετά κομψά- τα επιμέρους κομμάτια, όχι μόνο παραμένουν άθικτα, αλλά αλληλεπιδρούν και με περιβάλλοντα παρόμοια με τα αρχικά, κάνοντας αντικείμενο -σε ένα πιο βασικό επίπεδο- τη δυνατότητα δημιουργίας πρωτεϊνών από "ανταλλακτικά": ***αν το μόριο δεν διπλώσει, οι σωστές κατανομές -αν και αναγκαίες- δεν επαρκούν***. Παράλληλα δημιουργείται μια μικρού μεγέθους, μονομερής, πλατφόρμα για παραπέρα πειράματα. Η αλληλουχία της είναι "φυσικού τύπου" (πχ. χωρίς εσωτερικές επαναλήψεις), δεν έχει κάποια λειτουργία που να περιπλέκει τα αποτελέσματα επόμενων πειραμάτων που θα γίνουν επάνω της, και προσφέρεται για σχεδιασμό οποιασδήποτε νέας λειτουργίας χωράει στο μεγεθος αυτό.

Η πρωτεΐνη ROP, έχει αποτελέσει στο παρελθόν τη βάση για μια σειρά μελετών, των επιπτώσεων που έχουν στη σταθερότητα μεταλλαγές, τόσο στον υδρόφοβο πυρήνα, όσο και στα συνδετικά τμήματα μεταξύ των ελίκων [Steif et al, 1995· Vlassi et al, 1994]. Ειδικά ο υδρόφοβος πυρήνας είναι ευαίσθητος ακόμη και σε -θεωρητικά- συντηρητικές σημειακές μεταλλαγές: υποκαθιστώντας κατάλοιπα λευκίνης του υδρόφοβου πυρήνα του μορίου με βαλίνη ή αλανίνη, ανά ένα ή σε συνδυασμό με κατάλοιπα γειτονικά στον χώρο (δημιουργώντας έτσι κενά στο πακετάρισμα, στον υδρόφοβο πυρήνα), η θερμοκρασία αποδιάταξης (T_m) του μορίου χαμήλωνε σημαντικά. Πειράματα σαν το παραπάνω είναι το επόμενο λογικό βήμα.

Στο πειραματικό μέρος, το τελικό μόριο είναι ένα (ολιγο)πεπτίδιο μήκους 62 αμινοξικών καταλοίπων, απλό στην δημιουργία του, αφού το γονίδιο, που κωδικεύει για την πρωτεΐνη ROP, είναι διαθέσιμο. Ακόμη και αν δεν υπήρχε το γονίδιο, αν και το (τελικό) μόριο είναι σχετικά μεγάλο για τις διαδικασίες σύνθεσης ολιγοπεπτιδίων, ένα γονίδιο (μήκους 186 βάσεων, συν τρεις για τον κωδικό λήξης) που να κωδικεύει για την αλληλουχία του, είναι περισσότερο από εφικτό. Την έκφραση του γονιδίου, θα μπορούσε να ακολουθήσει μια πρώτη χρωματογραφική εκτίμηση του μεγέθους του προϊόντος μορίου στη φυσική του κατάσταση, και της διαλυτότητάς του. Επιπλέον, η μοναδική στο υπό σχεδιασμό μόριο τρυπτοφάνη, αν το μόριο διπλωθεί σωστά, κρύβεται στο εσωτερικό· αυτό αλλάζει τις φασματοσκοπικές της ιδιότητες, πράγμα που μπορεί να χρησιμεύσει σαν μια ακόμη ένδειξη ότι το μόριο διπλώθηκε όπως σχεδιάστηκε. Αυτά τα δεδομένα μπορούν να συλλεχτούν γρήγορα, και αν αποδειχτούν σύμφωνα με τα αναμενόμενα, τότε το ποσοστό της α-έλικας στη δομή μπορεί να προσδιοριστεί φασματοσκοπικά, ενώ το μόριο είναι αρκετά μικρό ώστε να μπορεί να προσδιοριστεί η δομή του και πειραματικά.

Αν και η τροποποίηση "φαίνεται" δραστική και θα μπορούσε να χαρακτηριστεί παρακινδυνευμένη (και περισσότερο το 1992 που σχεδιάστηκε), πολύ αργότερα, πραγματοποιήθηκε ένα παρόμοιο πείραμα από τους MacBeath et al [1998], όπου "μονομερίστηκε" ένα (διμερές στη φύση) ένζυμο. Οι εκτενείς ομοιότητες αφορούν τόσο το διμερισμό, που γίνεται μέσω ενός δεματιού, στο οποίο κάθε μονομερές συνεισφέρει δυο έλικες, όσο και στη σχέση μεταξύ των α-ελίκων ανά δύο με δυαδικό άξονα συμμετρίας. Οι ερευνητές αυτοί, επίσης, εκμεταλλεύτηκαν τη εσωτερική συμμετρία και ανήγαγαν -επίσης- το πείραμα σε δημιουργία ενός κατάλληλου συνδετικού τμήματος στο μονομερές, ώστε όλα τα απαιτούμενα μέρη του τελικού (μισού) μορίου να προέρχονται από ένα μόνο μονομερές. Το εγχείρημα όμως βασίστηκε ιδιαίτερα σε δοκιμή-και-σφάλμα (τυχαία μετάλλαξη και επιλογή πιο συγκεκριμένα) στο πειραματικό μέρος

Αντίθετα -και πρέπει να τονιστεί- ***στην παρούσα διαδικασία, που είχε σαν αποτέλεσμα το σχεδιασμό μιας τόσο δραστικής αλλαγής, με πολλαπλούς στόχους, βοήθησε ιδιαίτερα η γνώση των κανόνων του δομικού προτύπου,*** έστω και στην απλή μορφή των στατιστικών δεδομένων κάθε τοπολογικά διακριτής θέσης: οποτεδήποτε υπήρχε αμφιβολία που αφορούσε την επιλογή της κατάλληλης συνιστώσας για ο,τιδήποτε (είδος συνδετικού τμήματος ανάμεσα στις α-έλικες, είδος αμινοξικού καταλοίπου για κάποια θέση, κ.α.), υπήρχε έτοιμη απάντηση σε κάποιο πίνακα. Έτσι, ο σχεδιασμός προχώρησε γρήγορα, με την αναγκαιότητα για εξέταση του μοντέλλου στην οθόνη γραφικών να περιορίζεται στις λεπτομέρειες της συγκεκριμένης δομής που δεν καλύπτονται από τους πίνακες και τα στατιστικά δεδομένα. Δηλαδή δεν χρειάστηκε να αναλωθούν ώρες λεπτομερούς εργασίας μπροστά στην οθόνη των γραφικών μόνο και μόνο για να επιλεγούν αμινοξικά κατάλοιπα κατάλληλα για διάφορες τοπολογικά διακριτές θέσεις, αφού στις περισσότερες περιπτώσεις ήταν αμέσως προφανή από τους πίνακες.

Φυσικά, η προσέγγιση, όπως παρουσιάστηκε εδώ, έχει -θεωρητικά- το μειονέκτημα ότι η αναζήτηση δεδομένων περιορίζεται στο ρεπερτόριο μιας μικρής ομάδας πρωτεϊνών. Όμως, αν η ομάδα αυτή είναι καλά ορισμένη από πλευράς κοινών χαρακτηριστικών μεταξύ των μελών, τότε τα δεδομένα που αφορούν οποιοδήποτε ερώτημα σε μια τέτοια διαδικασία, βρίσκονται με μεγάλη πιθανότητα στη συλλογή αυτή. Σε αυτό το πλαίσιο, ελέγχθηκε μήπως κάποιο *άλλο* συνδετικό τμήμα, από οποιαδήποτε πρωτεΐνη (και με οποιαδήποτε γεωμετρία), μπορούσε να αντικαταστήσει είτε το ένα είτε το άλλο συνδετικό τμήμα στο μοντέλλο. Χρησιμοποιώντας την αντίστοιχη δυνατότητα, από το πρόγραμμα SYBYL, αναζητήθηκαν στην εσωτερική βάση δεδομένων του προγράμματος, τμήματα πρωτεϊνών που να συνδέουν την 15I με την 18'T και την 42'H με την 43D (σε αρίθμηση σύμφωνη με της πρωτεΐνης ROP), με δεδομένα το απαιτούμενο μήκος των (υπό σχεδιασμό) συνδετικών τμημάτων και τα εκατέρωθέν τους στη δομή τμήματα (σαν "αγκυροβόλια", *anchors*), χωρίς να προκύψουν αξιόλογες εναλλακτικές.

> Για την περίπτωση του τμήματος QADNP μόνο συνδετικά τμήματα με την ίδια γεωμετρία μπορούσαν να ενώσουν απροβλημάτιστα τα δύο μισά, ενώ για το άλλο συνδετικό τμήμα όλες οι

προτεινόμενες λύσεις περιελάμβαναν εισδοχές ή απαλοιφές καταλοίπων, οι δε τελευταίες αφαιρούσαν μια θέση d (που αντιστοιχεί στην 87W(2HMZ), Εικόνα 3), μειώνοντας ιδιαίτερα τον αριθμό υδρόφοβων αλληλεπιδράσεων στο υπό σχεδιασμό μόριο.

Η επιλογή τμημάτων πρωτεϊνών, από βάσεις δομικών δεδομένων, σε διαδικασίες δημιουργίας τρισδιάστατων μοντέλλων πρωτεϊνών, όπως μόλις περιγράφηκε, δηλαδή σε βοηθητικό ρόλο κατά το σχεδιασμό (συνήθως εντοπισμένων) μεταλλαγών για πειράματα κατευθυνόμενης μεταλλαξογένεσης, ήταν ίσως το μόνο σημείο σε διαδικασίες σχεδιασμού της εποχής, που είχε χροιά από δημιουργία πρωτεϊνών από ανταλλακτικά. Εξάλλου, δοθέντων των θέσεων (στο χώρο) των καταλοίπων, που πρέπει να ενωθούν με ένα συνδετικό τμήμα κάποιου μήκους, μπορεί να χρησιμοποιηθεί μια διαδικασία όπως των Bruccoleri και Karplus [1987]: ξεκινώντας από το ένα ελεύθερο άκρο, παράγει in silico όλες τις πιθανές διαμορφώσεις για ένα μικρό τμήμα και ελέγχει ποιες φτάνουν στο άλλο, με σωστή γεωμετρία, και αλληλεπιδρώντας όσο γίνεται ευνοϊκά με το υπόλοιπο μόριο. Όμως, μια αναζήτηση του είδους, για τμήματα μεγαλύτερα από λίγα (πχ 5) κατάλοιπα, επιστρέφει πολλές εναλλακτικές με μικρή ενεργειακή διαφορά μεταξύ τους. Προηγούμενη ανάλυση του δομικού προτύπου μπορεί να προσφέρει τα κριτήρια με τα οποία να επιλεγεί -με μεγαλύτερη πιθανότητα- η σωστή.

Ιστορικά, αρχικά (Φθινόπωρο 1992) είχα προτείνει ένα άλλο πείραμα, πάλι ξεκινώντας από την πρωτεΐνη ROP, με στόχο πάντα κάποια δραστική αλλαγή, πάλι στην κατεύθυνση του ελέγχου της επάρκειας ενός ελαττωμένου υδρόφοβου πυρήνα. Συγκεκριμένα, το σχέδιο προέβλεπε την αφαίρεση των κεντρικών επτάδων των α-ελίκων της πρωτεΐνης ROP, κάτι που θα άφηνε την πρωτεΐνη με έξι φέτες αλληλεπιδράσεων, δηλαδή δύο λιγότερες (αυτές που δείχνονται στην Εικόνα 7/Γεν. Εισαγωγή). Θα μπορούσε να υλοποιηθεί σαν ένα ολιγοπεπτίδιο μήκους 42 αμινοξικών καταλοίπων (56 που συμμετέχουν στο δεμάτι μείον 14 των δύο επτάδων), που θα δίπλωνε με τον τρόπο ίδιο με την πρωτεΐνη ROP φυσικού τύπου, δηλαδή σαν ομοδιμερές οργανωμένο γύρω από έναν άξονα συμμετρίας (Εικόνα 1α). Αν αυτό ήταν όντως το αποτέλεσμα, θα έδινε μια ακόμη ισχυρή ένδειξη για την ανεξαρτησία μεταξύ των διαδοχικών επτάδων και την επάρκειά τους. Η περίπτωση της διαίρεσης μιας πρωτεΐνης στα δύο, που περιγράφηκε λεπτομερώς παραπάνω, αν και πιο δραστική (αφαιρεί τέσσερις φέτες από τις οκτώ), είναι πιο κατάλληλη για λόγους επίδειξης, αφού περιλαμβάνει παρεμβάσεις που χρησιμοποιούν όλων των ειδών τους πίνακες και τα δεδομένα της παρούσας εργασίας. Αξίζει δε να σημειωθεί ότι το μόριο που προκύπτει *είναι μοντέλλο* (αν και καμιά συνιστώσα του δεν σχεδιάστηκε από το μηδέν), στο βαθμό που -χάρη σε άθικτα επιμέρους κομμάτια *και μεταξύ τους σχέσεις*- απομονώνει ένα συγκεκριμένο ερώτημα προς απάντηση.

Αναφέρθηκε με το κλείσιμο της εισαγωγής του κεφαλαίου, ότι πίνακες σαν εκείνους των Κεφ.Α.Ι και ΙΙ, μπορούν να διευκολύνουν όλο το φάσμα των δραστηριοτήτων σχεδιασμού. Έχοντας περιγράψει αναλυτικά ένα παράδειγμα, μπορεί να ολοκληρωθεί πιο άνετα η συζήτηση σχετικά με τις διάφορες επιπλέον δυνατότητες, αν και με τον κίνδυνο να αναφερθούν κάποια πράγματα, που πλέον θα πρέπει να είναι αυτονόητα.

Προκειμένου για μια σημειακή παρέμβαση, όπου κάποια θέση μεταλλάσσεται ώστε να διερευνηθεί η λειτουργική της σημασία, οι πίνακες αυτοί δείχνουν ποιοι άλλοι αμινοξικοί τύποι *ενδέχεται* να είναι συμβατοί, στα πλαίσια πάντα της στατιστικής φύσης της προτίμησης και της σχετικής σύστασης. Κατευθύνουν το πείραμα στα πλέον υποσχόμενα μονοπάτια, αφού για παράδειγμα, αν ένα φορτισμένο κατάλοιπο, βρίσκεται σε μια θέση, όπου κατά τεκμήριο (με βάση την ανάλυση του προτύπου) αναμένονται υδρόφοβα, τότε αποτελεί πρωταρχικό στόχο για ανάλυση. Επιπλέον, χρησιμεύουν στην ερμηνεία των εξαγομένων αφού μειώνουν την ασάφεια που επιφέρει η δομική παράμετρος της λειτουργικότητας μιας θέσης (δηλαδή, κάνουν λιγότερο αβέβαιο αν η λειτουργία του μορίου μειώθηκε επειδή η θέση που μεταλλάχτηκε έχασε τη διαμόρφωσή της). Το φορτίο, στο ίδιο παράδειγμα, καθώς μπορεί να αποτελεί θέση πρόσδεσης μετάλλου, ή να απαιτεί άλλο ένα φορτίο για σχηματισμό δεσμού άλατος[28], μπορεί να απαιτεί διαφορετικές (ποιοτικά) εξηγήσεις για τη συμπεριφορά του στις μεταλλαγές. Εξάλλου, αν σε κάποιο συνδετικό τμήμα, μια γλυκίνη σε διαμόρφωση α_L μεταλλαγεί σε προλίνη, παύει να μπορεί να έχει αυτή τη διαμόρφωση, οδηγώντας -ενδεχόμενα- ακόμη και σε ένα μόριο που δεν διπλώνει, κάτι που με γνώση των πιθανών διαμορφώσεων για συνδετικά τμήματα αυτού του μήκους σε αυτό το πρότυπο ίσως να μπορεί να διερευνηθεί από πριν. Βέβαια, απαιτείται προηγούμενη γνώση της δομής· όμως αυτή δεν χρειάζεται να είναι λεπτομερειακή, αφού συχνά μόνο η τοπολογία του μορίου αρκεί (“τι είναι μέσα”, “τι είναι έξω”, ή “τι είναι στις άκρες μιας α-έλικας”). Μάλιστα, οι ίδιοι πίνακες μπορούν να χρησιμεύσουν στη δημιουργία ενός ακριβέστερου τρισδιάστατου μοντέλλου, με βάση πρωτεΐνες οριακά ομόλογες.

> Φυσικά, αν πρόκειται να μεταλλαγεί μια μόνο θέση σε όλους τους τύπους, η διαδικασία της εκ των προτέρων ανάλυσης του δομικού προτύπου μοιάζει με τη χρήση μιας μηχανής τραίνου για να συρθεί το καρότσι ενός βρέφους (αν και δεν θα είναι πρόβλημα, αν -όπως αναφέρθηκε στην εισαγωγή του κεφαλαίου- υπάρχουν όλα σε μια τράπεζα πληροφοριών). Αν όμως επίκειται συστηματική διερεύνηση πολλών θέσεων, με ταυτόχρονες προσθαφαιρέσεις κομματιών, τότε μια τέτοια ανάλυση αναμένεται πολύ χρήσιμη.

[28] Δεσμοί άλατος κρυμμένοι στο εσωτερικό, συχνά σχετίζονται με την ειδικότητα στο δίπλωμα.

Προφανής είναι η βοήθεια, σε όλα τα στάδια, και κατά το σχεδιασμό ενός ολιγοπεπτιδίου, που προορίζεται να ολιγομεριστεί, δίνοντας στο σύνολο ένα μόριο-μοντέλλο. Όταν δε ο απώτερος στόχος είναι η συνένωση των ολιγοπεπτιδίων αυτών σε μια αλυσίδα, προσφέρονται έτοιμα τα είδη των δυνατών συνδέσεων και τα χαρακτηριστικά της αμινοξικής αλληλουχίας, με την οποία θα υλοποιηθεί. Αφού επιλεγεί ο "σκελετός" του μορίου (προκειμένου για δεμάτι: βαθμός ολιγομερισμού (2 ή 4), μήκος α-ελίκων, σε συνάρτηση με την δεξιό/αριστερόστροφη συνδεσμολογία, και τις ακτινικές θέσεις a-g, που αυτή ενδεχόμενα επιβάλλει στα σχετικά N/C-άκρα, αποστάσεις μεταξύ των α-ελίκων), για πολλές θέσεις ο κατάλληλος αμινοξικός τύπος είναι απ'ευθείας αναγνώσιμος από τους πίνακες. Για τις υπόλοιπες, τα δεδομένα των πινάκων μπορούν να συνδυαστούν με "*απευθείας"* υπολογιστικές τεχνικές (direct ή brute-force methods, όπως αποκαλούνται), που γίνονται πλέον εφικτές με την διαρκή πτώση του κόστους της υπολογιστικής ισχύος. Χαρακτηριστικό παράδειγμα τέτοιας, καθαρά συνδυαστικής, προσέγγισης είναι η αναζήτηση της κατάλληλης αλληλουχίας, μέσα από 1.9×10^{27} πιθανές, που διπλώνει σαν zinc finger, αλλά χωρίς ψευδάργυρο [Dahiyat και Mayo, 1997]. Αφού ορίστηκαν οι αμινοξικοί τύποι που επιτρέπονται σε κάθε θέση, δοκιμάστηκαν όλοι οι πιθανοί συνδυασμοί αμινοξικών τύπων στις διάφορες θέσεις, σε όλες τις πιθανές διαμορφώσεις των πλευρικών αλυσίδων τους. Εκτός από τους ισχυρότερους επεξεργαστές, βελτίωση θα πρέπει να αναμένεται και στους αλγορίθμους. Για παράδειγμα, μια τεχνική, που -αρχικά- προορίζεται για την εκτίμηση αλλαγών στη σταθερότητα μιας πρωτεΐνης από σημειακές μεταλλαγές, η ιδιοσύμφωνη βελτιστοποίηση συλλογής (self consistent ensemble optimisation, SCEO· Lee, 1994), είναι πολλά υποσχόμενη για τον από πριν υπολογισμό του πακεταρίσματος των πλευρικών αλυσίδων ανάμεσα στα στοιχεία της κύριας αλυσίδας. Ξεκινώντας από μια κατανομή των πλευρικών αλυσίδων στις διάφορες διαμορφώσεις, με διάφορες πιθανότητες, σταδιακά μειώνονται τα ποσοστά για εκείνες που αντιμετωπίζουν κάποιο πρόβλημα, και αυξάνουν για τις υπόλοιπες, μέχρι το σύστημα να ηρεμήσει σε μια κατάσταση ισορροπίας. ***Σε τέτοιες προσεγγίσεις, πίνακες σαν εκείνους των προηγουμένων κεφαλαίων, προσφέρουν το "πεδίο ορισμού",*** δηλαδή στη μεν πρώτη τους προτιμητέους αμινοξικούς τύπους για κάθε θέση, ενώ στη δεύτερη τις προτιμητέες αρχικές διαμορφώσεις. Εξάλλου, τέτοια προγράμματα θα πρέπει να συλλαμβάνουν ατέλειες, όπως η αντικατάσταση 16R(ROP)->D της παρούσης, όπου ένα θετικό φορτίο αντικαθίσταται με ένα αρνητικό, και να προτείνουν διορθώσεις. Έτσι, πληροφορία τέτοιου είδους, οπλισμένη με τέτοιες τεχνικές, για τις λεπτομέρειες που δεν περιλαμβάνονται άμεσα στους πίνακες, θα πρέπει να σταδιακά να μεταβάλλει σε ρουτίνα, εργασίες όπως εκείνες των ομάδων των DeGrado, Eisenberg και συνεργατών, οι οποίες αναλύονται σε κάποια λεπτομέρεια, προκειμένου να συζητηθούν κάποια επιπλέον σημεία.

> Πιο συγκεκριμένα, σε μια σειρά εργασιών [DeGrado et al, 1989], δοκιμάστηκε η φιλοσοφία ότι οι πρωτεΐνες μπορούν να θεωρηθούν σαν ένα σύνολο από μικρότερες ολιγοπεπτιδικές αλυσίδες ομοιοπολικά συνδεδεμένες· η καθεμία από αυτές δεν έχει συγκεκριμένη δομή από μόνη της, αλλά λειτουργώντας συνεργιστικά όλες μαζί οργανώνονται στην συγκεκριμένη τελική τρισδιάστατη δομή, οδηγημένες από τις μεταξύ τους (μεγάλης εμβελείας) αλληλεπιδράσεις. Έτσι, ξεκίνησαν από μικρά πεπτίδια σχεδιασμένα να

υιοθετούν δομή α-έλικας, εμπνευσμένα από τα μοντέλλα της τροπομυοσίνης των Hodges και συνεργατών [Lau et al, 1984], και φτάνοντας σε μια σειρά από σχεδιασμένα *de novo* τελικά προϊόντα.

Σαν πρώτο βήμα, σχεδίασαν ένα απλό πεπτίδιο, με αλληλουχία GELEELLKKLKELLKG, και με στόχο, σε τετραμερή μορφή, ένα δεμάτι από τέσσερις αντιπαράλληλες α-έλικες· το μόνο υδρόφοβο κατάλοιπο στον υπό σχεδιασμό υδρόφοβο πυρήνα είναι η λευκίνη, ενώ γλουταμικό και λυσίνη καταλαμβάνουν τις υπόλοιπες θέσεις. Σε μεγάλη αραίωση, το πεπτίδιο αυτό δεν παρουσιάζει καθόλου β'-ταγή δομή, ενώ σε αυξημένη συγκέντρωση η μετάβαση από μονομερή σε τετραμερή έδειχνε έναν έντονο συνεργιστικό χαρακτήρα. Σε ένα επόμενο βήμα ένωσαν δύο α-έλικες ομοιοπολικά με την αλληλουχία Pro-Arg-Arg δημιουργώντας μια φουρκέττα που σε διμερή κατάσταση θα έδινε το δεμάτι. Τέλος, ένωσαν δύο φουρκέττες με την ίδια συνδετική αλληλουχία, σε ένα μόριο που σε μονομερή μορφή θα δίπλωνε σε ένα πλήρες δεμάτι.

Φασματοσκοπία κυκλικού διχρωισμού έδειξε ότι το μόριο έχει υψηλό περιεχόμενο σε α-έλικα, ενώ το μέγεθος στη φυσική κατάσταση βρέθηκε να είναι το εκείνο της μονομερούς μορφής. Το σημαντικότερο είναι ότι, σε πειράματα αποδιάταξης με GuHCl η καμπύλη αποδιάταξης έδειξε ένα σημείο μετάπτωσης στην κατάσταση αποδιάταξης μεταξύ 6 και 7M, ενώ οι περισσότερες φυσικές πρωτεΐνες αποδιατάσσονται μεταξύ 1 και 3M. Η συγκέντρωση αυτή αντιστοιχεί σε μια διαφορά ενέργειας μεταξύ διπλωμένης δομής και κατάστασης αποδιάταξης της τάξης των 22 Kcal/mol. Άρα -από μια πρακτική σκοπιά- ***όχι μόνο μπορεί να σχεδιάσει κανείς πρωτεΐνες χωρίς να είναι ακόμη κατανοητή η διαδικασία του διπλώματος, αλλά μπορούν να είναι και εξαιρετικά σταθερές***. *Η επιτυχία της σταδιακής και σε μικρά βήματα διαδικασίας, που επέτρεψε να βελτιστοποιηθούν χωριστά οι έλικες και χωριστά τα συνδετικά τμήματα, μπορεί εν μέρει να είναι και ενδεικτική του πως σε μια πρωτεΐνη πρέπει τόσο να ικανοποιούνται οι ανάγκες κάθε συνιστώσας χωριστά, όσο και να ταιριάζουν αρμονικά μεταξύ τους όπου επικαλύπτονται.* Μπορεί επίσης να ακολουθεί τα πρώτα βήματα της εξέλιξης από μικρά πεπτίδια προς τις δομικές ενότητες των πρωτεϊνών.

Όμως, περισσότερο ενδιαφέρον είχε (όπως εξελίχθηκε) ένα πεπτίδιο, που αρχικά σχεδιάστηκε για τον έλεγχο της συμβατότητας κάθε αμινοξικού τύπου με τη δομή της α-έλικας [O'Neil και DeGrado, 1990], και τελικά είχε ένα απρόσμενο αποτέλεσμα. Το πεπτίδιο σχεδιάστηκε σαν τέσσερις επαναλήψεις της βασικής επτάδας LEALEGK με κάποιες τροποποιήσεις, και με στόχο ένα δεμάτι από δύο παράλληλες α-έλικες (όπως της τροπομυοσίνης). Οι θέσεις υπ' αριθ. 14, που είχαν σχεδιαστεί εξωτερικές στο δεμάτι αυτό, αλλάχτηκαν συστηματικά σε όλους τους αμινοξικούς τύπους, τα δε προϊόντα πήραν ονόματα του τύπου Coil-X, όπου X το κατάλοιπο που είχε τοποθετηθεί στις θέσεις 14. Όταν όμως η δομή της Coil-Ser προσδιορίστηκε, βρέθηκε να σχηματίζει ένα δεμάτι τριών α-ελίκων, με τις δύο παράλληλες και τη μια αντιπαράλληλη. Ενεργειακή ανάλυση έδειξε ότι οι διαφορές, μεταξύ της διευθέτησης αυτής και άλλων εναλλακτικών, είναι μικρές, κάτι που μας επιστρέφει στη συζήτηση ότι ***η αλληλουχία των πρωτεϊνών δεν κωδικεύει απλά για τη σωστή δομή, αλλά και μόνο γι'αυτήν, αποκλείοντας παρόμοιες αλλά λανθασμένες βιολογικά εναλλακτικές***. Παρόμοιο ήταν το πρόβλημα, σε μια προσπάθεια (με λιγότερο σαφή έκβαση) για πρωτεϊνικά μεμβρανικά κανάλια [Lear et al, 1988], αποτελουμένα από τέσσερις ή έξι (ασύνδετες) α-έλικες με μήκος 21 αμινοξικά κατάλοιπα, πακεταρισμένες σαν παράλληλο δεμάτι και με τις απλές αλληλουχίες $(LSLLLSL)_3$ και $(LSSLLSL)_3$, όπου τα κατάλοιπα σερίνης θα πρέπει να δείχνουν προς το εσωτερικό και τα κατάλοιπα λευκίνης στον υδρόφοβο περίγυρο. Αν και προκαταρκτικά δεδομένα έδειξαν υψηλό περιεχόμενο σε α-έλικα και αλλαγές στην αγωγιμότητα των μεμβρανών που χρησιμοποιήθηκαν, περιπλοκές από γεγονότα όπως οι πολλαπλές πιθανές καταστάσεις ενσωμάτωσης στη μεμβράνη ή ολιγομερισμού εμπόδισαν να δοθεί μια τελική απάντηση κατά πόσο το αποτέλεσμα συμφωνεί με το κανάλι που σχεδιάστηκε.

Εκτός από τις πολύ απλές αλληλουχίες, πρέπει να προσεχτεί το γεγονός ότι τέσσερις χωριστές α-έλικες που οργανώνονται σε ένα τετραμερές (έστω το σωστό), αναμένεται να έχουν άλλη κινητική και θερμοδυναμική συμπεριφορά από μια ενιαία αλυσίδα. Απλουστευμένες αλληλουχίες έχουν χρησιμοποιηθεί στο παρελθόν και για θεωρητικές αναλύσεις. Για παράδειγμα, οι Carlacci και Chou [1990α,β], σε μια προσπάθεια να εκτιμήσουν τη σχετική συνεισφορά α-ελίκων και συνδετικών τμημάτων στη σταθερότητα του δεματιού, βασίστηκαν σε ένα μοντέλο αποτελούμενο από τέσσερις α-έλικες πολυ-αλανίνης. Μάλιστα αρχικά, καθώς δεν είχαν συνυπολογίσει τα συνδετικά τμήματα, το δεμάτι που προέκυψε είχε ρομβοειδή αντί για τετράγωνη διατομή, κάτι που διορθώθηκε όταν αυτά συμπεριλήφθηκαν. Τα αποτελέσματα της παρούσης δείχνουν πόσο απέχει από οποιοδήποτε ρεαλισμό ένα τέτοιο μοντέλο, αφού ήδη α-έλικες πολυ-λευκίνης θα ήταν πολύ καλύτερες. Κατόπιν επισήμανσής μας [Paliakasis και Kokkinidis, 1991], επανέλαβαν τους υπολογισμούς με πραγματικά μόρια [Chou, 1991· Chou και Zheng, 1992], αλλά ποτέ με αλληλουχίες μοντέλλα, που θα είχαν όλα τα χαρακτηριστικά του δεματιού, χωρίς τις ιδιαιτερότητες καμίας συγκεκριμένης οικογένειας.

Εξάλλου, η δυνατότητα δημιουργίας συγκεκριμένων και δυνητικά χρήσιμων προϊόντων με νέες ιδιότητες, στη βάση τέτοιων πινάκων, έχει επίσης δειχτεί και μάλιστα με αλληλουχίες αρκετά σύνθετες και πολύπλοκες όπως των πρωτεϊνών που συναντώνται στη φύση. Συγκεκριμένα πρόκειται για μια πρωτεΐνη σχεδιασμένη να είναι πλούσια σε τέσσερα από τα βασικά[29] (για την διατροφή) αμινοξέα [Beauregard et al, 1995· Parker και Hefford, 1997]. Καθώς οι συγκεκριμένοι αμινοξικοί τύποι ευνοούν το σχηματισμό α-ελίκων, διάλεξαν το δεμάτι σαν υποψήφιο δομικό πρότυπο. Σε αντίθεση με μια προηγούμενη προσπάθεια της ομάδας των Richardson [Hecht et al, 1990], η διαδικασία σχεδιασμού βασίστηκε σε μια λεπτομερή ανάλυση του δομικού προτύπου. *Χρησιμοποιώντας σαν δείγμα τέσσερις από τις πρωτεΐνες του δείγματος της παρούσας*, έλεγξαν τις φυσικοχημικές ιδιότητες που θα έπρεπε να έχουν τοπολογικά διακριτές θέσεις κατά μήκος της αλληλουχίας. Έτσι σχεδίασαν, εξέφρασαν και χαρακτήρισαν κατ'αρχήν μια πρωτεΐνη μήκους 100 αμινοξικών καταλοίπων (11 Kd). Αυτή η προσέγγιση δείχνει την άμεση βοήθεια που μπορούν να προσφέρουν τέτοιου είδους αναλύσεις στο σχεδιασμό χρήσιμων, εφαρμοσμένων προϊόντων.

> Οι Hecht et al [1990] είχαν επιχειρήσει την κατασκευή μιας πρωτεΐνης, που επίσης θα υιοθετούσε τη δομή του δεματιού. Χρησιμοποιήθηκαν απλές αρχές δόμησης πρωτεϊνών: ο στόχος ήταν ένα δεμάτι από τέσσερις αντιπαράλληλες α-έλικες, μήκους πέντε περίπου στροφών, αποτελούμενες από κατάλοιπα που ευνοούν τις α-έλικες, με τη σωστή εναλλαγή στην υδροφοβικότητα μεν, αλλά χωρίς εσωτερικές επαναλήψεις στην αλληλουχία, και χωρίς ομολογία με οποιαδήποτε γνωστή πρωτεΐνη. Επιπλέον, δόθηκε προσοχή στις στατιστικές των άκρων των α-ελίκων, που ήταν ήδη γνωστές [Richardson και Richardson, 1988], και σε στοιχειώδεις απαιτήσεις των συνδετικών τμημάτων. Τέλος, πάρθηκαν μια σειρά προφυλάξεις, ώστε η πρωτεΐνη να μην μπορεί να ακολουθήσει άλλους παρόμοιους τρόπους διπλώματος. Για παράδειγμα, οι υδρόφοβες πλευρές δύο συνεχόμενων α-ελίκων δεν μπορούσαν να αποτελέσουν μια συνεχόμενη ταινία, αν η αλληλουχία τους σχημάτιζε μια μακριά α-έλικα. Για να παραχθεί τελικά, η μήκους 79 αμινοξικών καταλοίπων πρωτεΐνη Felix (όπως ονομάστηκε), φτιάχτηκε ένα συνθετικό γονίδιο, πλούσιο σε κωδικώνια που ευνοούν την έκφραση σε *E. coli*. Επειδή το προϊόν χάνοταν από εκτενή πρωτεόλυση, έγινε υπερέκφραση, με αποτέλεσμα σωμάτια εγκλεισμού (*inclusion bodies*), χάρη στα οποία αποφεύγονταν η πρωτεόλυση, και από τα οποία απομονώθηκε η πρωτεΐνη. Βιοχημικά και φασματοσκοπικά δεδομένα, στη συνέχεια, έδειξαν ότι πρόκειται για μια μονομερή πρωτεΐνη στο διάλυμα, αποτελούμενη κυρίως από α-έλικα, ενώ η μοναδική τρυπτοφάνη του μορίου κρύβεται στο υδρόφοβο εσωτερικό. Το εγχείρημα πέτυχε με την πρώτη προσπάθεια, όμως η όλη διαδικασία σχεδιασμού διήρκεσε πολύ περισσότερο.

Διαδικασίες σαν τις παραπάνω, πέρα από τη σημασία τους στην κατανόηση της δόμησης των πρωτεϊνών, δείχνουν ότι ο σχεδιασμός *συγκεκριμένων προϊόντων* δεν μπορεί να βασίζεται σε γενικές αρχές, αλλά σε πληροφορία ειδική κατά περίπτωση, ενώ θα πρέπει να περιλαμβάνει σαφείς τρόπους ελέγχου τους στο κάθε βήμα. Η επάρκεια της πληροφορίας (ή/και των κανόνων), που χρησιμοποιείται,

[29] Η ιδέα πίσω από την MB1 ήταν για ένα γονίδιο, που θα κωδικοποιούσε για μια πρωτεΐνη πλούσια σε βασικά αμινοξέα, εκείνα δηλαδή που ο οργανισμός πρέπει να παραλάβει από την τροφή, γιατί δεν μπορεί να τα συνθέσει· αυτή η πρωτεΐνη εκφρασμένη από τη βακτηριακή χλωρίδα στο σώμα των βοοειδών θα μπορούσε να ανεβάσει την ποιότητα των πρωτεϊνών του

είναι το πρώτο που δοκιμάζεται σε κάθε πείραμα που ακολουθεί: όσο πιο συγκεκριμένο είναι το ζητούμενο τόσο καλύτερης ευκρίνειας πληροφορία απαιτείται. Επιπλέον, προκειμένου για προϊόντα με πιθανές εφαρμογές ευρείας κλίμακας, θα πρέπει να είναι σαφές από την αρχή, πως θα γίνει κλιμάκωση κατά την παραγωγή. Όσο η κατανόηση των αρχών της δόμησης των πρωτεϊνών είναι ελλειπής, το μεγαλύτερο μέρος της προσπάθειας αναλώνεται στη δημιουργία της πλατφόρμας που θα φιλοξενήσει την επιθυμητή ιδιότητα. Όμως, με την πρόοδο της σχετικής γνώσης, η κατάσταση αυτή θα πρέπει να αλλάξει, και το βάρος σε αυτές τις διαδικασίες να επικεντρώνεται στις επιθυμητές ιδιότητες καθ'εαυτές: καθώς η de novo σύνθεση μιας πρωτεΐνης, που απλά διπλώνει σε μια προκαθορισμένη δομή, θα γίνεται όλο και περισσότερο ρουτίνα στο κοντινό μέλλον, η επιτυχία τέτοιων πειραμάτων θα κρίνεται όλο και περισσότερο από την χρησιμότητα των προϊόντων τους. Η διαδικασία που ακολουθήθηκε εδώ, όσον αφορά το σχεδιασμό της πλατφόρμας καθ'εαυτή, αφήνει πολύ λίγα σημεία για απάντηση από την πειραματική υλοποίηση: πολύ σύντομα, το αρχικό πλαίσιο εργασίας δίνει τη θέση του σε διαδικασίες αναζήτησης σε βάσεις δεδομένων, που συμπληρώνουν τα κενά του πλαισίου αυτού (έτσι επιλέγησαν τα συνδετικά τμήματα στη δική μας περίπτωση, αν και η "βάση" ήταν το δείγμα μας), ενώ ενεργειακοί (πχ. προσωμοίωση μοριακής δυναμικής ή άλλοι θεωρητικοί) υπολογισμοί αναλαμβάνουν την εξατομίκευση του προβλήματος, με πλήρη εκμετάλλευση ενός ισοζυγίου μεταξύ πληροφοριών τόσο γενικής φύσεως και προέλευσης (π.χ. στατιστικές), όσο και ειδικής για το συγκεκριμένο πρόβλημα και αποκαλύπτουν τα πιθανά ευαίσθητα σημεία του προϊόντος· διαρκής αλληλεπίδραση με τον ανθρώπινο παράγοντα (πχ. μέσα από οθόνες γραφικών) μπορεί να βοηθά στην επιλογή του επόμενου βήματος. Έτσι, όσον αφορά το σχεδιασμό της πλατφόρμας, μένει μόνο κάποια βελτιστοποίηση στη φάση της πειραματικής υλοποίησης (πχ με τυχαία μεταλλαξογένεση εντοπισμένη στα πιθανά αδύνατα σημεία, σε συνδυασμό με κάποιο πρωτόκολλο επιλογής), ενώ η εστία, όσον αφορά την έκβαση του πειράματος στο εργαστήριο, μετατοπίζεται στην απάντηση συγκεκριμένων ερωτημάτων.

Τα δομικά πρότυπα, που απαντώνται στη φύση, είναι λίγα ή επειδή δεν μπορούν να γίνουν άλλα ή επειδή η φύση μετέφερε λίγες επιτυχημένες λύσεις εξελικτικά. Τι ισχύει, και αν μπορούν να δημιουργηθούν περισσότερα, είναι ένα από τα πιο γοητευτικά θέματα στο χώρο της δομής των πρωτεϊνών. Στο μεταξύ όμως, ο μικρός αριθμός τους σημαίνει, ότι μπορούμε να εκμεταλλευόμαστε αναλύσεις του είδους, με χρήσιμους και αποτελεσματικούς τρόπους.

γάλακτος, που λόγω προέλευσης είναι πτωχή στα βασικά αμινοξέα.

Μέρος Β / Κεφάλαιο ΙΙ:

Αντί Επιλόγου: Που είναι η Πληροφορία;

Εισαγωγή

Μετά τη παρεμβολή της πιο άμεσης ίσως εφαρμογής των αναλύσεων του είδους, η συζήτηση επιστρέφει στα ερωτήματα που έμειναν προσωρινά χωρίς απάντηση στα προηγούμενα κεφάλαια. Έτσι, η εργασία αυτή ολοκληρώνεται με εκτιμήσεις της στατιστικής σταθερότητας των πινάκων του "Μέρους Α" καθώς και των δυνατοτήτων τους για πρόβλεψη δομής από αλληλουχία. Συζητείται η διαφορά των πινάκων αυτών από αντίστοιχους ενός συναφούς προτύπου, δίνοντας ίσως μια ένδειξη πως η φύση καταφέρνει αυτό που αναφέρθηκε πολλές φορές, δηλαδή να κωδικεύει όχι μόνο για τη σωστή δομή, αλλά και ενάντια σε λανθασμένες εναλλακτικές. Η όλη διερεύνηση, που πραγματοποιήθηκε από το Φθινόπωρο'94 μέχρι το Καλοκαίρι'95, αποκαλύπτει την πραγματική φύση τους, και τις επιπτώσεις σε ένα μοριακό εξελικτικό επίπεδο. Όλα αυτά μαζί συνθέτουν την απάντηση του αρχικού ερωτήματος, σε ποια μορφή βρίσκεται η πληροφορία για τη δομή μιας πρωτεΐνης κατά μήκος της αλληλουχίας της.

Στατιστική σταθερότητα

Όταν ο αριθμός των ανεξάρτητων παρατηρήσεων, σε ένα δείγμα, είναι (ή έστω φαίνεται) μικρός, καλό είναι να εξετάζεται η αξιοπιστία των συμπερασμάτων, που προκύπτουν από αυτό, η πιθανότητα δηλαδή με την οποία, η χρήση ενός άλλου δείγματος, παρομοίου μεγέθους και από τον ίδιο πληθυσμό, θα οδηγούσε σε διαφορετικά συμπεράσματα. Αυτό συνήθως γίνεται χρησιμοποιώντας εκτιμήτριες της σταθερότητας των μέσων όρων που υπολογίζονται, όπως η τυπική απόκλιση, που είναι ένα μέτρο της διασποράς των τιμών, από τις οποίες προέκυψε ο μέσος όρος. Όμως, ορισμένες από τις εκτιμήτριες έχουν νόημα όταν η κατανομή των τιμών (του υπό ανάλυση μεγέθους) ακολουθεί κάποιο συγκεκριμένο πρότυπο (π.χ. Gauss-Laplace), ενώ ο υπολογισμός τους δεν είναι πάντα, ούτε άμεσος, ούτε απλός -και για τους πίνακες των Κεφ. A.I-II *δεν είναι*, λόγω της κλιμακωτής εξαγωγής των αποτελεσμάτων (δηλαδή, πρώτα μέσα σε οικογένειες και έπειτα μεταξύ των οικογενειών). Μια εναλλακτική, που χρησιμο-ποιείται συχνά, είναι η διαίρεση του δείγματος σε μικρότερα μέρη, και επανάληψη της διαδικασίας για κάθε τμήμα χωριστά, οπότε, από την κατανομή των μέσων όρων, μπορούν να υπολογιστούν τα διαστήματα εμπιστοσύνης, που δείχνουν με ποια πιθανότητα ο μέσος όρος ενός μεγέθους κινείται μεταξύ κάποιων (άνω και κάτω) ορίων. Όμως, διαιρώντας το δείγμα των Κεφ. A.I-II έστω σε δύο μέρη, ήδη το κάθε επιμέρους δείγμα έχει μόνο 3-4 οικογένειες.

Στη δημοσίευση που αφορά το Κεφ. A.I [Paliakasis και Kokkinidis, 1992], έγινε μια ατυχής προσπάθεια να συμπεριληφθεί μια τυπική απόκλιση. Συγκεκριμένα, για κάθε τιμή, υπολογίστηκε η τυπική απόκλιση για τις επτά οικογένειες, π.χ. για τη σχετική σύσταση:

$$SD(\mathbf{O}_{(i,k,f)}) = \sqrt{\{\Sigma_{f=1,7}[\mathbf{O}_{f(i,k)} - \langle\mathbf{O}_{(i,k)}\rangle]^2 / 7\}}$$

όπου $\langle\mathbf{O}_{(i,k)}\rangle$ είναι ο μέσος όρος από τις επτά οικογένειες (f=1-7), για τον αμινοξικό τύπο i=1-20 στην θέση k=1-7 (δηλαδή a-g), όπως παρουσιάστηκαν στον Πίνακα A.I.4. Αυτή η τυπική απόκλιση δείχνει την διακύμανση της κάθε τιμής (π.χ. του ποσοστού των θέσεων d που καταλαμβάνονται από λευκίνη) από οικογένεια σε οικογένεια, κάτι που έχει μηχανιστική και όχι στατιστική βάση: η αύξηση του ποσοστού των αρωματικών στη φερριτίνη οφείλεται στις ιδιαίτερα αυξημένες αποστάσεις μεταξύ των α-ελίκων και δεν σχετίζεται με στατιστικής φύσεως παρέκκλιση· επίσης, στις αιμερυθρίνες, μια λιγότερο δραματική αύξηση των αποστάσεων μεταξύ των α-ελίκων, οδηγεί σε αύξηση του ποσοστού της λευκίνης σε θέσεις a. Ούτε δείχνει πάντα τέτοια "συνεργιστική" συμπεριφορά, αφού π.χ. συχνά η απόσταση μεταξύ δύο α-ελίκων είναι πολύ διαφορετική στο ένα άκρο σε σχέση με το άλλο. Δεν δείχνει την διακύμανση της πιθανότητας να βρεθεί ένας συγκεκριμένος αμινοξικός τύπος σε μια από τις επτά θέσεις, πολύ περισσότερο δεν δείχνει αν η αυξημένη ή ελαττωμένη παρουσία ενός αμινοξικού τύπου μπορεί να είναι τυχαία, οφειλόμενη σε στατιστική παρέκκλιση.

Πιο σωστή, και καθώς τα συμπεράσματα προκύπτουν κυρίως από τους αμινοξικούς τύπους με έντονη παρουσία, είναι η πρώτη πρόχειρη εκτίμηση της πιθανής διακύμανσης του ποσοστού της λευκίνης σε θέση d, που δόθηκε στο Κεφ. A.I: ακόμη και με μια μόνο αλληλουχία στην κάθε οικογένεια, θα είχαν καταμετρηθεί 80 θέσεις d, για τις οποίες, "31% λευκίνη" σημαίνει ότι οι 25

-περίπου- καταλαμβάνονται από λευκίνη. *Υποθέτοντας* ότι η κατανομή προσεγγίζεται από την Poisson στο κυρίως σώμα της (η ουρά εδώ δεν ενδιαφέρει), η τυπική απόκλιση είναι 5 κατάλοιπα. Το Κεφ. Α.Ι έμεινε με αυτή την πρόχειρη ένδειξη, γιατί ήταν -ιστορικά- η πρώτη που σκέφτηκα μετά τις τυπικές αποκλίσεις, που -ατυχώς- είχαν ήδη συμπεριληφθεί στη σχετική δημοσίευση. Όμως, ούτε αυτή δείχνει αν η αυξημένη ή ελαττωμένη παρουσία ενός αμινοξικού τύπου μπορεί να οφείλεται στη τύχη.

Ενδεικτικότερο είναι το (φαινομενικά παραπλήσιο αλλά τελείως διαφορετικό στην ουσία του) σκεπτικό που χρησιμοποιήθηκε στο Κεφ. Α.ΙΙ, το οποίο εδώ προσαρμόζεται για την περίπτωση της λευκίνης. Ας υποτεθεί ότι, στο γενικό πληθυσμό των δεματιών, η λευκίνη απαντά σε ένα ποσοστό των θέσεων d, συγκρίσιμο με εκείνο του καταλοίπου στις πρωτεΐνες (είτε γενικά, είτε εκείνες που αποτελούνται κυρίως από α-έλικα), δηλαδή 9%, και ότι αποκλίνει στο συγκεκριμένο δείγμα.

> ***Τεχνική σημείωση 1:*** Δεν μπορεί να χρησιμοποιηθεί η σύσταση του δεματιού σε λευκίνη, αφού επηρεάζεται σε μεγάλο βαθμό από την εξεταζόμενη ως αποκλίνουσα σύσταση της θέσης d.
>
> ***Τεχνική σημείωση 2:*** Υπενθυμίζεται ότι για n ανεξάρτητες δοκιμές, με πιθανότητα επιτυχίας p η καθεμία, ο αριθμός των (αναμενόμενων) επιτυχιών ακολουθεί τη δυωνυμική κατανομή (Bernoulli), δηλαδή η πιθανότητα να βρεθούν (ακριβώς) k επιτυχίες είναι $\boldsymbol{P(k,n:p)} = [\boldsymbol{n}! / (\boldsymbol{k}! \cdot (\boldsymbol{n-k})!)]\cdot\boldsymbol{p}^{\boldsymbol{k}} \cdot(1-\boldsymbol{p})^{\boldsymbol{n-k}}$. Η κατανομή αυτή, προσεγγίζεται από την κανονική, με μέσο όρο n·p και τυπική απόκλιση ίση προς √[n·p·(1-p)], με το σφάλμα κυρίως στις εκατέρωθεν "ουρές" (παρά για το κυρίως σώμα της κατανομής), όχι όμως ιδιαίτερα για n·p>(περίπου)10.

Θα αναμένονταν, λοιπόν, 80x9%=7,2 κατάλοιπα λευκίνης (σε θέση d), με τυπική απόκλιση μόλις 2,56 κατάλοιπα. Οπότε, τα 25 αντιστοιχούν σε μια τιμή 7 (περίπου) τυπικές αποκλίσεις μακριά από τον αναμενόμενο μέσο όρο. Ακόμη και αν ληφθεί υπ'όψη ότι η θέση d είναι κρυμμένη στο εσωτερικό, και αναμένεται να καταλαμβάνεται από υδρόφοβα κατάλοιπα, ανεβάζοντας (τηρουμένων των αναλογιών για τη λευκίνη) το ποσοστό σε 18%, θα αναμένονταν 14,4 κατάλοιπα λευκίνης, με τυπική απόκλιση 3,436· θα υπήρχε λοιπόν λιγότερο από 1‰ πιθανότητα να παρατηρηθούν 25 κατάλοιπα λευκίνης (μετά από αναγωγή σε πιθανότητα, μέσω της προσεγγίσεως της δυωνυμικής από την κανονική, των 3,08 τυπικών αποκλίσεων, στις οποίες αντιστοιχεί η διαφορά). Φυσικά, αντίστοιχα ισχύουν και για τη λευκίνη σε θέση a, ενώ η περίπτωση της αλανίνης συζητείται παρακάτω.

Αυτό το σκεπτικό, είναι τελείως διαφορετικό στην ουσία του: *εξετάζει το αποτέλεσμα σε σύγκριση με αυτό που θα αναμένοταν* (με βάση γενικούς μέσους όρους), και βοηθά στην ερμηνεία, δίνοντας τη ζητούμενη πιθανότητα μια παρατήρηση να είναι τυχαία, ενώ δεν χρειάζεται να γίνει καμιά υπόθεση για την κατανομή (*πυκνότητα*) της υποκείμενης πιθανότητας, προκειμένου να υπολογιστούν ο αναμενόμενος μέσος όρος και η τυπική απόκλιση του αριθμού των "επιτυχιών" -εκτός της

ανεξαρτησίας των 80 θέσεων d του "ανηγμένου" δείγματος. Αντίθετα, το προηγούμενο σκεπτικό *εξετάζει τη διασπορά των μετρούμενων τιμών*, οπότε έπρεπε να υποτεθεί ότι ο αριθμός παρατηρήσεων (ενός αμινοξικού τύπου σε μια θέση) ακολουθεί κάποια κατανομή (πχ την Poisson), και -τελικά- δεν εξηγεί τίποτε.

Επίσης, η σημαντικότητα των συμπερασμάτων ενισχύεται από την λογικά ομοιόμορφη -και εξηγήσιμη- συμπεριφορά αμινοξικών τύπων με παρόμοιες ιδιότητες. Ειδικά για τα υδρόφιλα κατάλοιπα σε εξωτερικές θέσεις, είναι προφανές ότι αν τα χειριστεί κανείς αθροιστικά (μιας και δεν έγινε τόσο έντονος διαχωρισμός για κάποια από αυτά, όσο για τη λευκίνη στο εσωτερικό), τα επίπεδα σημαντικότητας θα είναι αντίστοιχα, αν όχι καλύτερα. Εξάλλου, εκείνο που ενδιαφέρει δεν είναι η σταθερότητα κάθε τιμής (από τις 140 του κάθε πίνακα) χωριστά, αλλά των συμπερασμάτων που προκύπτουν από το σύνολο του κάθε πίνακα: αν, για κάθε τιμή, υπάρχει μια μικρή πιθανότητα, να αλλάξει περισσότερο από κάποια "ανεκτά" όρια, τότε το μικρό ποσοστό των τιμών, που θα μεταβληθούν, δεν πρόκειται να μεταβάλλουν την ουσία συμπερασμάτων που βασίζεται σε ένα σύνολο τιμών.

Όμως, καθώς από τον αρχικό υπολογισμό των πινάκων των Κεφ. A.I-II το 1990 μέχρι το 1995, δημοσιεύτηκαν οι πειραματικά προσδιορισμένες δομές πολλών ακόμη πρωτεϊνών, που υιοθετούν το δεμάτι σαν μέρος ή σύνολο της γ'-ταγούς δομής τους, παρέχοντας ένα περίπου τριπλασίου μεγέθους δείγμα, η ανάλυση -κατόπιν απαιτήσεως της τριμελούς συμβουλευτικής επιτροπής- επαναλήφθηκε. Τα στοιχεία δίνονται εν συντομία στον Πίνακα 1, αφού σκοπός δεν είναι η παράθεση ενός ακόμη Κεφ. A.I, αλλά η πιστοποίηση της στατιστικής σταθερότητας. Σε αντίθεση με το δείγμα των Κεφ. A.I-II, που δημιουργήθηκε με βάση τη βιβλιογραφία, εδώ η συλλογή του δείγματος έγινε με εκτενή χρήση ηλεκτρονικού υπολογιστή.

Συγκεκριμένα, η επιλογή έγινε με τη βοήθεια μιας σειράς προγραμμάτων, που το καθένα επέλεγε (ή απέρριπτε) μια πρωτεΐνη, με βάση ένα χαρακτηριστικό, περιορίζοντας σταδιακά τη συλλογή. Επελέγησαν από την Protein Data Bank [PDB, έκδοση Απρ. 1995· Bernstein et al, 1977] όσες πρωτεΐνες είχαν τέσσερα τμήματα (μήκους τουλάχιστο 6 καταλοίπων το καθένα) σε διαμόρφωση α-έλικας (για την ακρίβεια: που να σχηματίζει η κύρια αλυσίδα του καθενός δύο διαδοχικούς δεσμούς υδρογόνου τύπου i->i+3 ή i->i+4, χωρίς να παρεμβάλλεται cis-πεπτίδιο), με άξονες αντι-παράλληλους (δες Κεφ. A.II για υπολογισμό τους), σε αποστάσεις μεταξύ 6.5 και 15E και με μεταξύ τους γωνίες μεγαλύτερες από 120^{O}.

Αυτή ήταν μια σειρά κριτηρίων από διάφορες που δοκιμάστηκαν· για παράδειγμα, ελέγχθηκε επίσης μήπως υπήρχαν κατάλληλα δεμάτια, με γωνίες μεταξύ των ελίκων από 90^{O} μέχρι 120^{O}. Σε περιπτώσεις δομών προσδιορισμένων κρυσταλλογραφικά, η

ασύμμετρη μονάδα -συχνά- δεν περιλαμβάνει ολόκληρο το λειτουργικό μόριο. Έτσι, κατά τις αναζητήσεις αυτές, με την εφαρμογή των κατάλληλων περιστροφών και μετατοπίσεων, δημιουργούνταν ολόκληρη η μοναδιαία κυψελίδα, ενώ εξετάστηκε η ύπαρξη δεματιού στα όρια με όποιες τυχόν από τις (26 στο σύνολο) γειτονικές στο χώρο μοναδιαίες κυψελίδες χρειάστηκε. Σε περιπτώσεις πολλαπλών καταχωρίσεων, επελέγησαν εκείνες με την καλύτερη διακριτικότητα. Αν και το όριο για την ανάλυση ήταν τα 3 Ε, εκείνες που τελικά επελέγησαν έχουν γενικά καλύτερες αναλύσεις. Όσες δομές επελέγησαν τελικά, εξετάστηκαν οπτικά για να βρεθούν ποιες ήσαν αποδεκτές σαν αντι-παράλληλα δεμάτια με συστροφή αριστερά.

Ο προσδιορισμός της σύστασης των επτά θέσεων (a-g) σε αμινοξικούς τύπους, στο νέο δείγμα, έγινε όπως στο Κεφ. Α.Ι· πάλι χρησιμοποιήθηκαν όσες αλληλουχίες μπορούσαν να συστοιχιστούν χωρίς ασάφειες μαζί με τις πρωτεΐνες γνωστής δομής, ενώ για τις αιμερυθρίνες τηρήθηκε η ομαδοποίηση των Κεφ. Α.Ι-ΙΙ.

Πίνακας 1.α. *Περιληπτικά το νέο δείγμα.*

Κωδ.PDB	Πρωτεΐνη
1BBH	Κυτόχρωμα c’
1BGC	G-CSF (granulocyte colony stimulating factor)
1CCD	Uteroglobin
1CPC	Φυκοκυανίνη c
1GMF	GM-CSF (granulocyte macrophage colony stimulating factor)
1LPE	Απολιποπρωτεΐνη Ε3
1RIB	Πρωτεΐνη R2 της αναγωγάσης των ριβονουκλεοτιδίων
1ROP	Πρωτεΐνη ROP
1TPL	Λυάση τυροσίνης-φαινόλης
256B	Κυτόχρωμα b562
2CCY	Κυτόχρωμα c’ (Συγγενικό, αλλά διάφορο του 1BBH)
2CPK	Πρωτεϊνική κινάση εξαρτώμενη από κυκλικό AMP
2HMZ 2MHR	Αιμερυθρίνη Μυοαιμερυθρίνη (όπως κεφάλαια Α.Ι-ΙΙ)
2INT	Ιντερλευκίνη 4
2SOD	Δισμουτάση υπεροξειδίου χαλκού-ψευδαργύρου
2TMV	Πρωτεΐνη καλύμματος του ιού TMV (tobacco mosaic virus - ιος μωσαϊκής του καπνού)
3HHR	Αυξητική ορμόνη (Σωματοτροπίνη)
3INK	Ιντερλευκίνη 2
3LZM	Λυσοζύμη του φάγου Τ4
	Φερριτίνη (όπως κεφάλαια Α.Ι-ΙΙ)

Πίνακας 1.β. *Σύσταση των επτά τοπολογικά διακριτών θέσεων του δεματιού σε αμινοξικούς τύπους με βάση το νέο διευρυμένο δείγμα. Πάνω η νέα τιμή, κάτω η παλιά (τιμές:%). Μ.Ο: μέσος όρος από τις επτά στήλες. Οι στήλες "Πρωτ" και "All-α" έχουν μεταφερθεί από τον Πίν. Α.Ι.5*

	a	b	c	d	e	f	g	M.O	All-α	Πρωτ.
A	11.3	9.2	9.6	11.5	15.2	13.4	11.6	11.7		
	15.0	7.0	11.0	19.0	12.0	13.0	17.0	13.4	11.6	8.7
C	3.9	1.0	1.9	1.9	0.3	0.6	0.9	1.5		
	6.0	0.0	2.0	3.0	1.0	0.0	1.0	1.9	1.7	1.6
D	2.8	7.9	7.6	0.9	1.9	7.1	4.5	4.7		
	1.0	12.0	8.0	0.0	4.0	12.0	3.0	5.9	6.5	5.7
E	3.1	11.4	10.0	1.4	6.8	12.1	5.4	7.2		
	5.0	8.0	16.0	0.0	5.0	9.0	7.0	7.2	6.5	6.4
F	6.8	2.6	2.1	6.8	2.3	1.4	5.1	3.9		
	10.0	2.0	4.0	7.0	3.0	1.0	2.0	4.2	4.2	3.9
G	2.2	2.9	6.3	1.4	2.3	7.1	3.9	3.7		
	1.0	4.0	2.0	1.0	3.0	2.0	6.0	2.7	7.7	7.8
H	2.6	1.5	2.3	3.0	3.9	3.0	2.7	2.7		
	3.0	1.0	2.0	5.0	6.0	3.0	4.0	3.1	2.8	2.1
I	7.0	1.1	2.5	9.5	6.9	1.8	4.7	4.8		
	9.0	2.0	2.0	6.0	9.0	3.0	3.0	4.6	3.7	5.1
K	1.6	13.9	7.8	0.6	9.8	9.6	3.9	6.7		
	1.0	16.0	8.0	1.0	6.0	11.0	6.0	7.0	10.1	6.8
L	26.2	5.5	5.7	26.6	12.1	3.2	13.2	13.2		
	20.0	8.0	2.0	31.0	17.0	1.0	8.0	12.2	8.9	9.2
M	7.1	1.6	1.2	3.4	3.5	0.3	7.5	3.5		
	6.0	1.0	0.0	5.0	3.0	0.0	5.0	3.0	2.4	2.1
N	0.5	6.3	8.0	0.9	3.4	6.3	7.9	4.8		
	0.0	7.0	10.0	0.0	3.0	11.0	10.0	5.7	3.8	4.4
P	1.6	2.4	1.0	0.9	1.9	1.8	0.9	1.5		
	0.0	1.0	1.0	1.0	3.0	3.0	0.0	1.3	3.8	4.5
Q	2.2	6.3	7.9	4.6	9.0	5.4	5.0	5.8		
	1.0	5.0	5.0	6.0	7.0	6.0	8.0	5.1	3.3	3.9
R	5.0	7.0	6.5	1.1	4.7	7.2	5.5	5.3		
	6.0	8.0	6.0	1.0	6.0	7.0	4.0	5.8	2.8	4.8
S	1.9	5.2	6.7	3.2	3.3	8.1	4.6	4.7		
	1.0	3.0	4.0	4.0	3.0	5.0	3.0	3.2	5.4	6.7
T	3.8	7.8	7.9	3.6	2.0	7.1	5.3	5.4		
	4.0	8.0	11.0	1.0	4.0	7.0	6.0	5.8	4.9	5.8
V	6.0	4.4	2.6	8.4	4.2	2.4	5.6	4.8		
	6.0	7.0	4.0	7.0	4.0	5.0	5.0	5.4	6.0	7.0
W	1.7	1.4	0.6	6.6	1.3	0.5	0.7	1.8		
	3.0	0.0	1.0	0.0	0.0	0.0	2.0	0.7	1.2	1.2
Y	2.9	0.6	1.6	3.6	5.4	1.6	1.2	2.4		
	2.0	0.0	2.0	3.0	1.0	1.0	2.0	1.7	2.6	3.3

Τα αποτελέσματα για την σύσταση των επτά θέσεων σε αμινοξικούς τύπους αντιπαραβάλλονται με τα αρχικά στον Πίνακα 1, όπου έχουν -επίσης- μεταφερθεί η μέση σύσταση των πρωτεϊνών (γενικά), η μέση σύσταση των πρωτεϊνών που αποτελούνται αποκλειστικά από α-έλικα και η μέση σύσταση του δεματιού από τον Πίνακα Α.Ι.5 για αντιπαραβολή με την καινούργια. Δεν υπάρχουν πλέον θέσεις κενές παρατηρήσεων ("μηδενικά"), ενώ συνοπτικά, παρατηρεί κανείς ότι τα βασικά συμπεράσματα για τη σύσταση των θέσεων δεν αλλάζουν. Ειδικά η μέση σύσταση ελάχιστα έχει αλλάξει, εκτός από το γεγονός ότι τώρα το δεμάτι φαίνεται πιο ανεκτικό σε γλυκίνη (αλλά κυρίως σε θέσεις c και f) και σερίνη. Κατά τα λοιπά, παραμένει το υψηλό ποσοστό σε λευκίνη, ενώ η αλανίνη, που η τιμή της ήταν κοντά στο μέσο όρο για πρωτεΐνες που αποτελούνται από α-έλικα, συγκλίνει ακόμη περισσότερο σε αυτόν.

Αντίθετα, στις επιμέρους μετρήσεις υπάρχουν διακυμάνσεις, συχνά "ανησυχητικές". Για περίπου 40% των τιμών, η απόκλιση είναι μέχρι περίπου 20% (είτε προς τα πάνω είτε προς τα κάτω), με περίπου το 70% από τα "υψηλά" ποσοστά που διαμόρφωσαν τα συμπεράσματα στα αντίστοιχα κεφάλαια στην ομάδα αυτή. Ίσως, απόκλιση της τάξης του 20% να φαίνεται -διαισθητικά- σημαντική για αυτά τα "μεγάλα" ποσοστά. Όμως, παραμένοντας στο παράδειγμα του ανηγμένου δείγματος των 80 θέσεων d, που αναφέρθηκε νωρίτερα, και σε στενή σύνδεση με αυτό, ας σημειωθεί ότι 20% για τις 25 λευκίνες (που αναμένονται σε αυτή τη θέση, με βάση το 31%), αντιστοιχεί σε 5 κατάλοιπα πάνω ή κάτω, που αναμένονται σαν τυπική απόκλιση, εύρος μέσα στο οποίο αναμένεται να κινείται το 67% ποσοστών τέτοιου μεγέθους. Εξάλλου, κάποιες από τις μεγάλες αποκλίσεις στα "υψηλά" ποσοστά απλά οδηγούν στην "εξομάλυνση" συμπερασμάτων που έχουν ήδη εξαχθεί. Χαρακτηριστικό παράδειγμα η αλανίνη, για την οποία αναφέρθηκε ότι (σε σχέση με τα οξύτατα μέγιστα άλλων αμινοξικών τύπων σε κάποιες θέσεις) παρουσιάζει μια πιο ομοιόμορφη συμμετοχή στις επτά θέσεις. Εδώ, για κάποιες θέσεις, οι τιμές συνέκλιναν προς το 11,5% από μεγαλύτερες (με μεταβολή εως 40%) ή μικρότερες τιμές (με μεταβολή εως 30%), φέρνοντας -επιπλέον- και τη μέση σύσταση του δεματιού για αλανίνη πιο κοντά στην αντίστοιχη τιμή της στήλης "All-α" του Πίνακα 1.β, κάνοντας -όπως γράφτηκε ήδη- το δεμάτι να διαφέρει ακόμη λιγότερο από τις πρωτεΐνες υψηλού ποσοστού σε α-έλικα, σε μια βασική του συνιστώσα. Άλλο παράδειγμα, η σύγκλιση των τιμών στις θέσεις b, c και f στο 7-8% για το ασπαρτικό (μεταβολές κατά 5%-40%), και στο 11% για το γλουταμικό οξύ (στις ίδιες θέσεις· μεταβολές κατά 30%-40%), που συμβάλλουν στην άρση της "ανησυχίας", που είχε διατυπωθεί στο κεφ. Α.Ι, μήπως (εκτός από τις εσωτερικές θέσεις a και d) δεν είναι ισοδύναμες μεταξύ τους ούτε οι εξωτερικές θέσεις. Αντίθετα, κάτι τέτοιο δεν μπορεί να λεχθεί για τις θέσεις e και g, παρά το γεγονός ότι από τις σημαντικές μεταβολές (≈50%) είναι η σύγκλιση στο 12% για τη λευκίνη στις θέσεις αυτές. Μένει επίσης να εξηγηθεί η άνοδος του ποσοστού της γλυκίνης σε κάποιες εξωτερικές θέσεις.

Είναι φανερό, ότι καθώς διαφορετικά συμπεράσματα απαιτούν απάντηση σε διαφορετικά ερωτήματα, η χρήση μιας μόνης εκτιμήτριας (πχ τυπική απόκλιση) δεν μπορεί έτσι κι αλλιώς να καλύψει τα πάντα. Π.χ., για την ομοιομορφία στην κατανομή

της αλανίνης, ας υποτεθεί ότι στο σύνολο των δειγμάτων (τέτοιων μεγεθών, 10-20 οικογένειες), που μπορούν να δημιουργηθούν από τον γενικό πληθυσμό, τρεις θέσεις (από τις επτά, όχι πάντα οι ίδιες) ξεφεύγουν από την ομοιομορφία αυτή, δηλαδή καταλαμβάνονται από αλανίνη σε ποσοστό πολύ επάνω ή κάτω από 11.5%. Η πιθανότητα, να βρεθεί δείγμα με όλες τις θέσεις κοντά στο 11.5%, είναι <2%.

Παρ'όλες τις διακυμάνσεις στις επιμέρους μετρήσεις, ποιοτικά, η σύσταση των επτά θέσεων δεν άλλαξε. Οι θέσεις εξακολουθούν να κυριαρχούνται από λίγους αμινοξικούς τύπους η καθεμία, ενώ και οι αμινοξικοί τύποι δεν άλλαξαν τις προτιμήσεις τους, σε γενικές γραμμές. Έτσι η θέση a εξακολουθεί να κυριαρχείται από λευκίνη και αλανίνη, ακολουθούμενα από τα υδρόφοβα φαινυλ-αλανίνη, ισολευκίνη, μεθειονίνη και βαλίνη (περίπου τα 2/3 των θέσεων αθροιστικά), ενώ η θέση d από λευκίνη και αλανίνη (αν και λιγότερο από ότι αρχικά) ακολουθούμενα από βαλίνη και φαινυλ-αλανίνη (περισσότερες από τις μισές θέσεις αθροιστικά). Τα φορτισμένα ασπαρτικό/γλουταμικό, αργινίνη, λυσίνη μαζί με τα πολικά ασπαραγίνη/γλουταμίνη, σερίνη και θρεονίνη κυριαρχούν στο εξωτερικό (3/4 των θέσεων b, c και f), με την αλανίνη να κρατά περίπου 11% (κατά μέσο όρο) από κάθε θέση και τη γλυκίνη να έχει ένα σχετικά αυξημένο ποσοστό. Τέλος, από τη γενική ασάφεια των θέσεων e και g, ξεχωρίζουν πάλι τα -υψηλά- ποσοστά της λευκίνης και της αλανίνης, με πολλά όμως άλλα κατάλοιπα να έχουν σημαντικά ποσοστά είτε στη μια θέση είτε και στις δυο. Μπορεί λοιπόν να γραφεί ότι, παρά τις αποκλίσεις στις επιμέρους τιμές, τα συμπεράσματα παρέμειναν σε γενικές γραμμές τα ίδια (που αποτελεί και το ζητούμενο).

Όμως, μήπως άλλαξε η αξιοπιστία και η σταθερότητα των παρατηρήσεων; και πόσο; Η απάντηση βρίσκεται στην αναλογία

<αναμενόμενη τυπική απόκλιση μιας μέτρησης>/<αναμενόμενη μέση τιμή μέτρησης>

η οποία για τη δυωνυμική είναι $\sqrt{[n\cdot p\cdot(1-p)]}/(n\cdot p) = \sqrt{[(1-p)/(n\cdot p)]}$ και για $1-p \cong 1$ (δηλαδή για τα μικρά ποσοστά) ισούται περίπου προς $1/\sqrt{(n\cdot p)}$. Δηλαδή, το μέγεθος της τυπικής απόκλισης, ως ποσοστό σε σχέση με τη μετρούμενη ποσότητα, ελαττώνεται, όχι ανάλογα με το μέγεθος της μετρούμενης ποσότητας, αλλά ανάλογα με την τετραγωνική του ρίζα. Αυτό σημαίνει ότι, παρά τον τριπλασιασμό του δείγματος, το μέγεθος αυτό μίκρυνε στο $\sqrt{[(1-p)/(3\cdot n\cdot p)]} / \sqrt{[(1-p)/(n\cdot p)]} = 1/\sqrt{3} \cong 58\%$ του αρχικού· δηλαδή, βελτιώθηκε κατά (περίπου) 40% , ανεξάρτητα από το αν πρόκειται για κάποιο από τα "υψηλά" ποσοστά ή όχι. Για μια μέτρηση μεγέθους 25 καταλοίπων, η τυπική απόκλιση ελαττώθηκε από τα 5 κατάλοιπα στα 3.

Διαφορές από συναφή πρότυπα

Εκτός από το δεμάτι, κι άλλα δομικά πρότυπα που αποτελούνται από α-έλικες μπορούν να περιγραφούν σε όρους επτάδων θέσεων a-g (ιδίως όπου οι έλικες υπερ-ελικώνονται και συστρέφονται η μια γύρω από την άλλη σαν ένα σκοινί). Συνήθως οι θέσεις που έρχονται σε επαφή με υδρόφοβο περιβάλλον ή απλά με τη διπλανή α-έλικα χαρακτηρίζονται σαν a και d, και οι υπόλοιπες ανάλογα. Γεννιέται λοιπόν το ερώτημα αν υπάρχει διαφορά στη στατιστική αυτών των επτάδων από του δεματιού. Οι Lupas και συνεργάτες [1991] έχουν διεξάγει μια ανάλυση, αντίστοιχη με την παρούσα, σε πρωτεΐνες που παρουσιάζουν διάφορες δομές υπερ-έλικας, από δύο ή τρεις παράλληλες α-έλικες. Συγκεκριμένα, μέτρησαν την σύσταση κάθε θέσης a-g στις δομικές ενότητες υπερ-έλικας των οικογενειών της τροπομυοσίνης, της μυοσίνης και της κερατίνης, και στη συνέχεια την διαίρεσαν με τη μέση σύσταση των πρωτεϊνών σε αμινοξικούς τύπους, υπολογίζοντας έτσι την *ανηγμένη σχετική σύσταση*.

Στον Πίνακα 2 αντιπαραβάλλονται τα αποτελέσματα των παραπάνω ερευνητών, με το ίδιο μέγεθος υπολογισμένο για το δεμάτι. Οι διαφορές που διαπιστώνονται είναι σημαντικές:

(α) Οι υπερ-έλικες της δικής τους ανάλυσης παρουσιάζουν πολλές περιπτώσεις ανηγμένης σχετικής σύστασης μεγαλύτερης από 2 (δηλαδή η σχετική σύσταση μιας θέσης σε κάποιο αμινοξικό τύπο να είναι υπερδιπλάσια από εκείνη των πρωτεϊνών για τον ίδιο τύπο), και μάλιστα για αμινοξικούς τύπους με υψηλά ποσοστά στη γενική σύσταση των πρωτεϊνών. Χαρακτηριστικότερη περίπτωση αποτελεί το γλουταμικό οξύ σε θέσεις b (3.5), c (3.1) και f (2.5) (αλλά και e και g, με 5.7 και 3.0, που συχνά είναι εκτεθειμένες σε αυτές τις δομές). Επίσης η λευκίνη σε θέσεις a και d (3.1 και 3.9), λυσίνη σε θέσεις b και g (2.6 και 2.8), και λιγότερο το ασπαρτικό οξύ· και επιπλέον η αλανίνη σε θέση d (2.6) όπως και γλουταμίνη / ασπαραγίνη σε κάποιες θέσεις με τιμές μεταξύ 2 και 2.5 .

(β) Υπάρχει χαμηλότερη ανεκτικότητα σε κατάλοιπα όχι ευνοϊκά για το πρότυπο: η προλίνη εξαφανίζεται, η σχετική σύσταση σε ιστιδίνη, κυστεΐνη και γλυκίνη είναι μικρότερη από του δεματιού, ενώ η ανοχή των θέσεων a και d για την ογκώδη τρυπτοφάνη είναι μειωμένη. Η σερίνη και η θρεονίνη διατηρούν τα ποσοστά που παρουσιάζουν στο δεμάτι.

(γ) Υπάρχουν αλλαγές στην ανηγμένη σχετική σύσταση για αρκετούς αμινοξικούς τύπους, με χαρακτηριστικά παραδείγματα την αλανίνη, και τα γλουταμικό/ασπαρτικό οξύ.

Πίνακας 2. *Επάνω τιμή (στο κάθε κελί): Ανηγμένη σχετική σύσταση των επτά θέσεων του δεματιού σε αμινοξικούς τύπους. Προκύπτει από τον Πίνακα 1.β, διαιρώντας τη σχετική σύσταση κάθε θέσης σε κάθε αμινοξικό τύπο με τη σύσταση των πρωτεϊνών στον ίδιο τύπο.*

Κάτω τιμή: Το ίδιο μέγεθος για τις επτά θέσεις στις υπερ-έλικες, όπως δίνονται από τους Lupas et al [1991]. Στη στήλη GB φαίνονται τα ποσοστά εμφάνισης των 20 αμινοξικών τύπων στις πρωτεΐνες (όπως προκύπτουν από καταμέτρηση στην αντίστοιχη έκδοση της βάσης δεδομένων GenBank), όπως δίνονται στην ίδια εργασία, και με βάση τα οποία έγινε η αναγωγή.

Τιμές άνω του 2.0 υπογραμμισμένες, άνω του 2.7 έντονες, και άνω του 3.0 με διπλή υπ/μιση.

	a	b	c	d	e	f	g	GB
A	1.30	1.06	1.10	1.32	1.75	1.54	1.33	
	1.30	1.55	1.08	2.61	0.38	1.25	0.88	7.59
C	2.44	0.63	1.19	1.19	0.19	0.38	0.56	
	0.82	0.02	0.31	0.15	0.18	0.16	0.04	1.86
D	0.49	1.39	1.33	0.16	0.33	1.25	0.79	
	0.03	2.35	2.27	0.24	0.66	1.62	1.45	5.03
E	0.48	1.78	1.56	0.22	1.06	1.89	0.84	
	0.26	**3.50**	**3.11**	1.00	**5.69**	2.49	**3.05**	6.10
F	1.74	0.67	0.54	1.74	0.59	0.36	1.31	
	0.53	0.08	0.40	0.66	0.19	0.11	0.01	3.88
G	0.28	0.37	0.81	0.18	0.29	0.91	0.50	
	0.05	0.28	0.58	0.22	0.21	0.43	0.16	7.10
H	1.24	0.71	1.10	1.43	1.86	1.43	1.29	
	0.35	0.28	0.68	0.40	0.29	0.58	0.21	2.25
I	1.37	0.22	0.49	1.86	1.35	0.35	0.92	
	2.60	0.10	0.35	0.89	0.51	0.47	0.43	5.35
K	0.24	2.04	1.15	0.09	1.44	1.41	0.57	
	1.38	2.64	1.76	0.19	1.82	1.96	**2.80**	5.72
L	**2.85**	0.60	0.62	**2.89**	1.32	0.35	1.43	
	3.17	0.30	0.40	**3.90**	0.59	0.50	0.48	9.33
M	**3.38**	0.76	0.57	1.62	1.67	0.14	**3.57**	
	2.24	0.37	0.48	1.41	0.54	0.77	0.66	2.34
N	0.11	1.43	1.82	0.20	0.77	1.43	1.80	
	0.84	1.48	1.53	0.04	1.72	2.46	2.28	4.25
P	0.36	0.53	0.22	0.20	0.42	0.40	0.20	
	---	0.01	---	0.01	---	---	---	5.28
Q	0.56	1.62	2.03	1.18	2.31	1.38	1.28	
	0.18	2.11	1.78	0.63	2.55	1.58	2.53	4.27
R	1.04	1.46	1.35	0.23	0.98	1.50	1.15	
	0.66	1.16	1.21	0.03	1.36	1.94	1.80	5.39
S	0.28	0.78	1.00	0.48	0.49	1.21	0.69	
	0.38	0.58	1.05	0.42	0.53	0.92	0.63	7.28
T	0.66	1.34	1.36	0.62	0.34	1.22	0.91	
	0.17	0.70	0.96	0.65	0.79	0.84	0.65	5.97
V	0.86	0.63	0.37	1.20	0.60	0.34	0.80	

	1.67	0.40	0.39	0.95	0.21	0.34	0.36	6.42
W	1.42	1.17	0.50	**5.50**	1.08	0.42	0.58	
	0.24	---	---	0.46	0.02	---	---	1.41
Y	0.88	0.18	0.48	1.09	1.64	0.48	0.36	
	1.42	0.09	0.12	1.66	0.19	0.13	0.16	3.16

(δ) Όμως, η σημαντικότερη διαφορά έρχεται εξετάζοντας τον πίνακα από τη σκοπιά των θέσεων. Πιο συγκεκριμένα, για ορισμένους αμινοξικούς τύπους που είναι χαρακτηριστικοί για κάποιες θέσεις, υπάρχουν σημαντικές διαφορές στη σχετική σύσταση:

(δ1) Η θέση d κατακυριαρχείται από λευκίνη και αλανίνη (αθροιστικά σε ποσοστά περίπου 40%+20%=60%, σε σύγκριση με 40% στο δεμάτι), ενώ η φαινυλ-αλανίνη, η ισολευκίνη και η βαλίνη, που στο δεμάτι συμπληρώνουν την εικόνα, κινούνται κάτω από το μέσο όρο τους για πρωτεΐνες. Στη θέση a, η λευκίνη και η αλανίνη κινούνται στα ίδια επίπεδα με το δεμάτι, όμως αυξάνει δραματικά (διπλασιάζεται) το ποσοστό των β-διακλαδισμένων βαλίνη και ισολευκίνη. Εξετάζοντας θέματα πακεταρίσματος της θέσης a στο δεμάτι (όπου "αντικρύζει" την γειτονική στο χώρο α-έλικα Εικόνα 7/Γεν.Εισαγωγή), διαπιστώνει κανείς ότι β-διακλάδωση δεν ευνοείται στη θέση αυτή, λόγω περιορισμών -που όμως δεν υπάρχουν στη θέση a των υπερ-ελίκων (το ένα γ-άτομο άνθρακα εκβάλλει στον κενό χώρο). Αντίθετα, η φαινυλ-αλανίνη, που συμπλήρωνε την κατανομή αυτής της θέσης στο δεμάτι, εδώ μειώνεται στο 1/3. Αυτό ίσως οφείλεται εν μέρει στην εντονότερη έκθεση του φαινολικού δακτυλίου από τη θέση a της υπερ-έλικας, και εν μέρει στο ότι η φαινυλ-αλανίνη στη θέση a του δεματιού συχνά απαντά σε ρόλο "καπακιού" του υδρόφοβου πυρήνα από την άκρη του δεματιού (δες Κεφ. Α.Ι), θέση που προφανώς στην υπερ-έλικα δεν υπάρχει.

(δ2) Στις θέσεις b, c και f, όπως αναφέρθηκε ήδη, παρατηρείται στην υπερ-έλικα μια αύξηση των ήδη υψηλών τιμών των πολικών καταλοίπων και κυρίως των φορτισμένων (με πρωταγωνιστή το γλουταμικό οξύ), σε σχέση με το δεμάτι. Όμως η δραματικότερη διαφορά παρατηρείται για τις θέσεις e και g. Στο δεμάτι, οι θέσεις αυτές βρίσκονται στο όριο, μεταξύ του υδρόφοβου πυρήνα και του εξωτερικού χώρου, ενώ στην υπερ-έλικα είναι εκτεθειμένες, με αποτέλεσμα να παρουσιάζουν μια κατακόρυφη αύξηση οι τιμές για τα πολικά και τα φορτισμένα, ενώ χάνονται τα υδρόφοβα κατάλοιπα. Έτσι, η ανηγμένη σχετική σύσταση σε γλουταμικό είναι 5.7 και 3.0 (από περίπου 1), σε λυσίνη 1.8 και 2.8 (από 1.4 και μόλις 0.6) και σε αργινίνη 1.4 και 1.8 (από περίπου 1)· ενώ αύξηση παρουσιάζει και στα πολικά ασπαραγίνη (1.7 και 2.3 από 0.8 και 1.8) και γλουταμίνη (2.5 και 2.5 από 2.3 και 1.3), πάντα για θέση e και g αντίστοιχα. Αντίθετα, τα υδρόφοβα λευκίνη, ισολευκίνη, βαλίνη, φαινυλ-αλανίνη, μεθειονίνη καθώς και η αλανίνη πέφτουν, στις περισσότερες περιπτώσεις, στο 0.2-0.5 της μέσης εμφάνισής τους στις πρωτεΐνες, από τιμές ≥ 1.

Αντίστοιχες παρατηρήσεις ισχύουν και για το σύνηθες πρότυπο διμερισμού "φερμουάρ λευκίνης" (leucine zipper). Σε αυτό το πρότυπο, δύο τμήματα α-έλικας, μήκους >22 κατάλοιπα πακετάρονται παράλληλα, και μπορούν επίσης να περιγραφούν μια επανάληψη επτά διαδοχικών θέσεων από τις οποίες οι a και d είναι μεταξύ των δύο ελίκων· μάλιστα κάθε θέση a είναι στο ίδιο ύψος με την αντίστοιχή της από την άλλη έλικα, όπως και κάθε θέση d, αντίστοιχα. Στα πλαίσια της παρούσης, έγινε μια πρόχειρη προσπάθεια για μια σχετική στατιστική ανάλυση (το καλοκαίρι του 1994 συγκεκριμένα). Η αναζήτηση βασίστηκε ιδιαίτερα στη σημείωση -από τους αντίστοιχους ερευνητές- μέσα στις ίδιες τις καταχωρήσεις στην τράπεζα δεδομένων (SwissProt [Bairoch και Boeckmann, 1991]), ότι περιέχεται ένα τέτοιο μοτίβο. Συγκεντρώθηκαν μερικές εκατοντάδες υπο-αλληλουχίες, με κύριο χαρακτηριστικό τις τέσσερις λευκίνες, 7 θέσεις μακριά η καθεμία από την επόμενη (στις θέσεις d). Το πρώτο αποτέλεσμα που προέκυψε ήταν ότι πάνω από 2/3 των θέσεων a καταλαμβάνονταν από βαλίνη, εκτός από τη θέση a μεταξύ των δύο "μεσαίων" λευκινών (από τις τέσσερις), που σε μεγάλο ποσοστό καταλαμβάνοταν από ασπαραγίνη. Η τρισδιάστατη δομή, για το μοτίβο αυτό, έχει προσδιοριστεί πειραματικά για την πρωτεΐνη GCN4 (κωδ. PDB: 2ZTA). Αυτό, το συνήθως άκρως αποσταθεροποιητικό για το μέσο μιας α-έλικας κατάλοιπο, εδώ οδηγεί σε δεσμό υδρογόνου μεταξύ των δύο α-ελίκων (η πλευρική αλυσίδα της ασπαραγίνης -σε θέση a- από τη μια έλικα, με την αντίστοιχη της άλλης έλικας, στο ίδιο ύψος). Τα αυξημένα ποσοστά, σε υδρόφιλα (ιδιαίτερα σε φορτισμένα) κατάλοιπα σε θέσεις e και g, παρατηρήθηκαν και εδώ. Οι λεπτομέρειες δεν έχουν σημασία, αφού η προσπάθεια εγκαταλήφθηκε, καθώς μεγάλο ποσοστό των υπο-αλληλουχιών δεν συμφωνούσε με τους κανόνες που προέκυπταν, και δεν ήταν καθαρό αν αυτό οφείλοταν σε όντως επιτρεπόμενες αποκλίσεις από τους κανόνες, ή οι ερευνητές είχαν εσφαλμένα καταχωρήσει σαν leucine zipper αλληλουχίες που δεν ήσαν. Ορισμένες από τις παραπάνω παρατηρήσεις, σχολιάζουν και οι DeGrado et al [1999], στα πλαίσια, όχι κάποιας στατιστικής ανάλυσης, αλλά ανασκόπησης διαδικασιών σχεδιασμού μικρών πεπτιδίων, σχεδιασμένων για τη διερεύνηση τέτοιων θεμάτων.

Η παρατήρηση (δ) -παραπάνω- *ίσως* σχετίζεται με *έναν* τρόπο που η φύση αποθήκευσε (εξελικτικά), όχι μόνο την πληροφορία για τη σωστή δομή, αλλά και την πληροφορία προς αποφυγή συναφών -καίτοι λανθασμένων- εναλλακτικών. Παρά το κοινό χαρακτηριστικό της λευκίνης και της αλανίνης στις θέσεις a και d, η μηχανή της ζωής δεν μπερδεύεται και δεν μετατρέπει με λίγες απλές (τυχαίες) μεταλλαγές μια αλληλουχία που διπλώνει με βάση το ένα πρότυπο στο άλλο, αφού η αλληλουχία που προορίζεται για αριστερόστροφο αντι-παράλληλο δεμάτι έχει πολλά χαρακτηριστικά

που δεν ευνοούν τη δημιουργία παράλληλης υπερ-έλικας, και αντίστροφα. Μάλιστα, αν σημειωθεί ότι η θέση a στο δεμάτι δεν είναι ισοδύναμη τοπολογικά με τη θέση a -π.χ.- στην υπερ-έλικα, ένα λάθος πακετάρισμα τέτοιων ελίκων, που σε επίπεδο αλληλουχιών περιγράφονται και οι δυο με επαναλήψεις επτάδας, καθίσταται ακόμη μεγαλύτερο λάθος.

Κλείνουμε το κομμάτι αυτό με μια παρατήρηση τεχνικής φύσεως, σχετικά με τον Πίνακα 2, πριν χρησιμοποιηθεί στα επόμενα. Καθώς, στο νέο δείγμα, ο αριθμός των οικογενειών τριπλασιάστηκε, ακόμη κι αν κάθε οικογένεια είχε μόνο έναν αντιπρόσωπο, θα υπήρχαν περίπου 250 παρατηρήσεις για την κάθε θέση (a-g). Είναι θεμιτό να χρησιμοποιηθεί αναγωγή σε ένα τέτοιο δείγμα, αφού όπως εξηγήθηκε από νωρίς (Κεφ. Α.Ι), οι επιπλέον αλληλουχίες μέσα στην κάθε οικογένεια, δεν αυξάνουν τον αριθμό των ανεξάρτητων παρατηρήσεων, αλλά απλά κάνουν πιο αξιόπιστες τις τιμές τους. Άρα, αν ένα κατάλοιπο επρόκειτο να εμφανιστεί με συχνότητα ανάλογη της γενικότερης εμφάνισής του στις πρωτεΐνες, και αυτή είναι της τάξης του 10%, αναμένονται 25 εμφανίσεις με τυπική απόκλιση περίπου 5 κατάλοιπα, και επομένως τιμές μεγαλύτερες του 36 (σχετική σύσταση=14,4%, ανηγμένη σχετική σύσταση=1,44) είναι στατιστικά σημαντικά διαφορετικές με πιθανότητα λάθους 1%. Αν η γενικότερη εμφάνισή του στις πρωτεΐνες είναι της τάξης του 6%, οι αναμενόμενες 15 εμφανίσεις έχουν τυπική απόκλιση περίπου 3,9 κατάλοιπα, και επομένως για πιθανότητα λάθους 1% στατιστικά σημαντικά διαφορετικές είναι τιμές μεγαλύτερες του 23,9 (σχετική σύσταση=9,5%, ανηγμένη σχετική σύσταση=1,6). Στον Πίνακα 2 έχουν υπογραμμιστεί τιμές ανηγμένης σχετικής σύστασης μεγαλύτερες από 2,0 -στατιστικά σημαντικά διαφορετικές, στο ίδιο όριο λάθους του 1%, ακόμη και για αμινοξικούς τύπους με ποσοστό γενικής εμφάνισης 2% (πχ κυστεΐνη). Ίσως πρέπει να διευκρινιστεί ο λόγος, που η διαίρεση πρέπει να είναι με τη μέση σύσταση των πρωτεϊνών σε κάθε τύπο, και όχι του δεματιού, όπως στον Πίνακα Α.ΙΙ.5 . Στον τελευταίο ενδιέφερε η προτίμηση ενός καταλοίπου να πάει σε συγκεκριμένες θέσεις εφ'όσον είναι ήδη στο δεμάτι. Εδώ -εκτός από το θέμα της σύγκρισης με την εργασία των Lupas et al [1991]- η συζήτηση αφορά τις διαφορές του δεματιού από έννοιες εκτός δεματιού (δηλαδή συναφή πρότυπα και τις πρωτεΐνες γενικά), ιδιαίτερα στα επόμενα όπου συζητείται η δυνατότητα εκτίμησης της συμβατότητας μιας αλληλουχίας με το δομικό πρότυπο του δεματιού.

Δυνατότητες για πρόβλεψη - Και μια έκπληξη

Το είδος της πρόβλεψης, που μπορούν να προσφέρουν τα δεδομένα μιας ανάλυσης, σχετίζεται στενά με το είδος των δεδομένων αυτών. Για παράδειγμα, στις "κλασσικές" στατιστικές μεθόδους πρόβλεψης, όπου αναλύεται η συμπεριφορά, είτε των αμινοξικών τύπων είτε μικρών τμημάτων αλληλουχίας, σε σχέση με τη συμμετοχή στα β'-ταγή στοιχεία σε πειραματικά γνωστές δομές, επιχειρείται έπειτα πρόβλεψη της θέσης των β'-ταγών στοιχείων μιας πρωτεΐνης με άγνωστη δομή. Η ανηγμένη σχετική σύσταση (Πίνακας 2) δείχνει πόσο περισσότερο (ή λιγότερο) έχει "ανάγκη" τον κάθε αμινοξικό τύπο η κάθε τοπολογικά διακριτή θέση, σε σχέση με τη μέση "ανάγκη" των πρωτεϊνών· κρίνοντας ένα μικρό τμήμα αλληλουχίας σύμφωνα με τις συνδυασμένες ανάγκες του συνόλου των θέσεων, μπορεί να συμπεράνει κανείς αν είναι κατάλληλο να συμμετάσχει σε ένα πιθανό δεμάτι. Σύμφωνα λοιπόν και με τη συζήτηση στο τέλος της Γενικής Εισαγωγής, η αντίστοιχη πρόβλεψη εμπίπτει στην κατηγορία της αναγνώρισης δομικού προτύπου.

Συνήθως η εξέταση της καταλληλότητας ενός τμήματος αλληλουχίας να υιοθετήσει ένα δομικό πρότυπο γίνεται με τον εξής τρόπο: Πολλαπλασιάζοντας μεταξύ τους τις τιμές της ανηγ. σχετικής σύστασης ν (πιθανολογουμένων ως) τοπολογικά διακριτών θέσεων στο υπό εξέταση τμήμα, για το αμινοξικό κατάλοιπο που θα βρεθεί στη κάθε θέση, και εξάγοντας την ν-οστή ρίζα του γινομένου, προκύπτει ο γεωμετρικός μέσος όρος. Αυτός θεωρείται ότι δείχνει την μέση "καταλληλότητα" (ανά κατάλοιπο) του τμήματος για να μετέχει σε δεμάτι. Φυσικά αυτό θα πρέπει να γίνει για όλα τα πιθανα πλαίσια ανάγνωσης (επτά στην περίπτωση του δεματιού), ή τις πιθανές αντιστοιχίσεις θέσεων στο τμήμα, στη γενικότερη περίπτωση. Οι Lupas και συνεργάτες [1991] χρησιμοποίησαν αυτό τον τρόπο για να δείξουν την ικανότητα των αποτελεσμάτων τους να διακρίνουν τμήματα αλληλουχίας κατάλληλα για υπερ-έλικες.

Για καθαρά τεχνικούς λόγους, συχνά μια ανηγμένη σχετική σύσταση μετατρέπεται σε λογαριθμική μορφή. Έτσι, όταν η επί τοις εκατό σύσταση μιας θέσης σε κάποιον αμινοξικό τύπο είναι ίδια με το μέσο όρο των πρωτεϊνών (τιμή στον Πίνακα 2 ίση με μονάδα), το αποτέλεσμα είναι μηδέν· όταν η θέση έχει περισσότερη "ανάγκη" τον συγκεκριμένο τύπο το αποτέλεσμα είναι θετικό· και όταν ο συγκεκριμένος τύπος είναι προβληματικός για τη θέση αυτή, η τιμή του λογαρίθμου είναι αρνητική. Απλά, στη λογαριθμική μορφή έχουμε άθροισμα αντί για γινόμενο μεταξύ των υποψηφίων θέσεων[30]. Ο Πίνακας 3 είναι η λογαριθμική μορφή του Πίνακα 2, πολλαπλασιασμένη επί 10, ώστε το αποτέλεσμα να είναι χονδρικά από -10 ως 10, και επομένως ευκολότερο να το παρακολουθήσει από την πλευρά του ο αναγνώστης.

Πίνακας 3. *Λογαριθμική μορφή των "επάνω" τιμών του Πίνακα 2. Προκύπτει πολλαπλασιάζοντας το λογάριθμο με βάση το 10 επί 10 (δηλαδή* $y=10*\log_{10}(x)$*).*

[30] Παρεμπιπτόντως, σε αυτή τη μορφή είναι συνήθως και οι πίνακες σύγκρισης/συστοίχισης μεταξύ αλληλουχιών.

	a	b	c	d	e	f	G
A	1.14	0.25	0.41	1.21	2.43	1.88	1.24
C	3.87	-2.01	0.76	0.76	-7.21	-4.20	-2.52
D	-3.10	1.43	1.24	-7.96	-4.81	0.97	-1.02
E	-3.19	2.50	1.93	-6.58	0.25	2.76	-0.76
F	2.41	-1.74	-2.68	2.41	-2.29	-4.44	1.17
G	-5.53	-4.32	-0.92	-7.45	-5.38	-0.41	-3.01
H	0.93	-1.49	0.41	1.55	2.70	1.55	1.11
I	1.37	-6.58	-3.10	2.70	1.30	-4.56	-0.36
K	-6.20	3.10	0.61	-10.46	1.58	1.49	-2.44
L	4.55	-2.22	-2.08	4.61	1.21	-4.56	1.55
M	5.29	-1.19	-2.44	2.10	2.23	-8.54	5.53
N	-9.59	1.55	2.60	-6.99	-1.14	1.55	2.55
P	-4.44	-2.76	-6.58	-6.99	-3.77	-3.98	-6.99
Q	-2.52	2.10	3.07	0.72	3.64	1.40	1.07
R	0.17	1.64	1.30	-6.38	-0.09	1.76	0.61
S	-5.53	-1.08	0.00	-3.19	-3.10	0.83	-1.61
T	-1.80	1.27	1.34	-2.08	-4.69	0.86	-0.41
V	-0.66	-2.01	-4.32	0.79	-2.22	-4.69	-0.97
W	1.52	0.68	-3.01	7.40	0.33	-3.77	-2.37
Y	-0.56	-7.45	-3.19	0.37	2.15	-3.19	-4.44

Ο αναμενόμενος μέσος όρος ανά κατάλοιπο του Πίνακα 3, σταθμίζοντας την κάθε τιμή με τη σχετική συχνότητα εμφάνισης του αντίστοιχου αμινοξικού τύπου στις πρωτεΐνες, είναι περίπου -1, με τυπική απόκλιση ίση προς 3,26 . Αυτό σημαίνει ότι, αν εκτιμηθεί η καταλληλότητα ενός τυχαίου τμήματος αλληλουχίας, με μήκος ν κατάλοιπα και με μέση σύσταση σε αμινοξέα ανάλογη με των πρωτεϊνών, με βάση τη διαδικασία που περιγράφτηκε παραπάνω, το αποτέλεσμα θα είναι περίπου -ν . Καθώς αθροίζονται τιμές από διαδοχικές θέσεις, αλλά ανεξάρτητα κατάλοιπα, η αναμενόμενη τυπική απόκλιση είναι $3{,}26\cdot\sqrt{\nu}$. Έτσι, αν εξετασθεί ένα τμήμα με μήκος 18 καταλοιπα (5 στροφές α-έλικας), όσο δηλαδή είναι το μέσο μήκος μιας α-έλικας ενός δεματιού, αναμένεται μια μέση βαθμολογία ίση προς -18 με αναμενόμενη τυπική απόκλιση ίση προς 13,8 .

Αν όμως η κάθε τιμή σταθμιστεί με τη σχετική συχνότητα εμφάνισης του αντίστοιχου αμινοξικού τύπου στην αντίστοιχη θέση, ο αναμενόμενος μέσος όρος ανά κατάλοιπο γίνεται περίπου +1 με τυπική απόκλιση περίπου 2,5 . Αυτό σημαίνει ότι αν δημιουργηθούν τμήματα όπου οι διαδοχικές θέσεις υπακούν στη σύσταση διαδοχικών θέσεων του δεματιού, το ίδιο τμήμα των 18 καταλοίπων αναμένεται τώρα να έχει μια βαθμολογία ίση προς +18 (και αναμενόμενη τυπική απόκλιση ίση προς 10,7), δηλαδή αναμένεται να ευρίσκεται περίπου 2,5 τυπικές αποκλίσεις μακριά από το μέσο όρο τμημάτων ίσου μήκους, αλλά με αλληλουχία τυχαιοποιημένη με βάση τη μέση σχετική σύσταση των πρωτεϊνών σε αμινοξέα. Δοκιμάζοντας α-έλικες πραγματικών δεματιών, με σκοπό την εύρεση τμημάτων με το τέτοιο μήκος στο κατάλληλο πλαίσιο ανάγνωσης, ώστε η βαθμολογία τους να είναι όσο το δυνατό περισσότερες τυπικές αποκλίσεις μακριά από το μέσο όρο των τυχαιοποιημένων, διαπιστώθηκε ότι κινούνται γύρω στις 3-4, αν και δεν έλειψαν περιπτώσεις που το καλύτερο τμήμα ήταν μόλις 2 τυπικές αποκλίσεις μακριά από αυτόν το μέσο όρο. Η δοκιμή ανήχθη σε μια απλή αλληλουχιακή συστοίχιση, όπου η μια αλληλουχία αποτελούταν από μια επανάληψη της διαδοχής $(abcdefg)_n$ και η άλλη ήταν η υπό εξέταση, με πίνακα βαθμολογίας τον Πίνακα 3, και χωρίς κενά στη συστοίχιση. Αν και -θεωρητικά- θα μπορούσε να χρησιμοποιηθεί οποιοσδήποτε αλγόριθμος συστοίχισης, ακολουθήθηκε εκείνος που περιγράφεται σε ξεχωριστή εργασία μετά την παρούσα, μιας και αποφεύγει προβλήματα στις προσεγγίσεις που ήσαν διαθέσιμες, και που περιγράφονται σε λεπτομέρεια στο αντίστοιχο κεφάλαιο. Σε συντομία, αυτός ο αλγόριθμος, αφού εντοπίσει μικρές "στοιχειώδεις" αντιστοιχίες μεταξύ μικρών τμημάτων (2-3 κατάλοιπα μήκος) των δύο αλληλουχιών, τις ενώνει *ανά δύο την κάθε φορά* σε όλο και μεγαλύτερες συστοιχίσεις, μέχρι το σημείο όπου παραπέρα συνενώσεις δεν βελτιώνουν τη συστοίχιση.

Καθώς το μήκος στο οποίο επιτυγχάνοταν η βέλτιστη διαφορά ήταν πάνω από 10 κατάλοιπα, η κατανομή της αθροιστικής βαθμολογίας μπορεί να προσεγγιστεί ικανοποιητικά από την κανονική· άρα, οι τυπικές αυτές αποκλίσεις μπορούν να μεταφραστούν σε πιθανότητες, μέσω της κανονικής, με ικανοποιητική ακρίβεια. Από αυτή τη "μετάφραση" προκύπτει ότι, λίγο λιγότερο από 1% των τμημάτων, αυτού του μήκους και με αλληλουχία τυχαιοποιημένη με βάση τη μέση σύσταση των πρωτεϊνών σε αμινοξέα, είναι εξίσου κατάλληλο για αριστερόστροφο αντι-παράλληλο δεμάτι, όσο είναι κατά μέσο όρο οι έλικες των πραγματικών δεματιών, ενώ 1‰ μπορεί να εγγίσει τα πλέον συμβατά με δεμάτι τμήματά τους. Αυτό δεν είναι τίποτε άλλο από την αριθμητική έκφραση εκείνου που γράφτηκε συχνά στα προηγούμενα, ότι οι περιορισμοί που θέτει ένα δομικό πρότυπο (τουλάχιστο το δεμάτι εν προκειμένω) είναι γενικής φύσεως και με μεγάλη ανοχή σε λάθη, όπου λέγοντας λάθη εννοούμε την ποικιλία αλληλουχιών που μπορούν να υιοθετήσουν το συγκεκριμένο πρότυπο.

Αυτή η παρατήρηση, όχι μόνο μπορεί να προκαλεί έκπληξη, αλλά θέτει και ένα δύσκολο πλαίσιο στο οποίο πρέπει να γίνονται εκτιμήσεις του είδους: το κάθε τμήμα μιας αλληλουχίας με άγνωστη δομή, δεν μπορεί να εκτιμηθεί χωριστά ως προς το αν είναι μέρος ενός δεματιού. Σε μια τράπεζα αλληλουχιών, μεγέθους γονιδιώματος (πχ 30000 αλληλουχίες, με συνολικά 10^7 κατάλοιπα), ακόμη και

απαγορεύοντας την αλληλεπικάλυψη των υπό εξέταση αλληλουχιών, μια τέτοια διαδικασία θα θεωρούσε κατάλληλες για δεμάτι 500-5000 υπο-αλληλουχίες (ανάλογα με την αυστηρότητα των κριτηρίων, και πάντως κατά πάσα πιθανότητα >1500, αφού συνήθως τέτοιες αναζητήσεις δεν γίνονται με τα αυστηρότερα), από εκείνες που *δεν* υιοθετούν το πρότυπο. Τηρώντας τις αναλογίες που συνάντησα κατά τη δημιουργία του διευρυμένου δείγματος το 1995, από τις 30000 αλληλουχίες, περίπου 1000 ακόμη αναμένονται να υιοθετούν το δεμάτι σαν δομικό πρότυπο, το 1/3 από τις οποίες δεν θα είναι ανιχνεύσιμες παρά μόνο με τα χαλαρότερα κριτήρια (εκείνα που θα δώσουν και 5000 λανθασμένες υποψήφιες!).

Έτσι, θα πρέπει να εκτιμάται συνολικά η αλληλουχία αν έχει όλα τα απαιτούμενα τμήματα (τέσσερα κατάλληλα για δεμάτι ή δύο, ανάλογα με το βαθμό ολιγομερισμού), που να τελειώνουν με τρόπο συμβατό με τα άκρα των ελίκων. Σε συνδυασμό και με τα αποτελέσματα του Κεφ. Α.ΙΙ, τα μεταξύ τους συνδετικά τμήματα θα πρέπει να ελέγχονται -ενδεχομένως- χωριστά, και -όταν είναι μικρά- θα πρέπει να ελέγχεται αν είναι συμβατά με πιθανές υποχρεωτικές διαμορφώσεις, όπως επίσης αν τελειώνουν στις σωστές ακτινικές θέσεις· ενώ σαφές θα πρέπει να είναι και το πως διευθετείται όση τυχόν αλληλουχία περισσεύει. Όπως για τα αμινοξικά κατάλοιπα αναφέρθηκε στην εισαγωγή, ότι οι στερεοχημικά επιτρεπτές διαμορφώσεις για κάθε κατάλοιπο είναι εκείνες που, αν εφαρμοστούν σε διαδοχικά κατάλοιπα, οδηγούν σε ικανοποίηση και των δυνατοτήτων της κύριας αλυσίδας για δεσμούς υδρογόνου, δίνοντας γένεση σε α-έλικες και β-κλώνους, έτσι και στο τριτοταγές επίπεδο, θα πρέπει σε ένα κοινό πλαίσιο να ικανοποιούνται όλες οι ανάγκες.

Η ανάγκη για μια τέτοια συνδυαστική προσέγγιση, προκύπτει από το ποσοστό της πληροφορίας που είναι αποθηκευμένη στην αλληλουχία σαν σύνολο, σύμφωνα με το κατανεμημένο πρότυπο. Το γεγονός ότι μπορούν να συνδυάζονται μεταξύ τους ολόκληρα τμήματα, αντανακλά το ποσοστό της πληροφορίας που είναι οργανωμένο κεντρικά, δηλαδή μέσα στο κάθε τμήμα και ανεξάρτητα από τα υπόλοιπα. Σε συνδυασμό δε και με τη συζήτηση (στη Γενική Εισαγωγή) περί ειδικότητας των αλληλεπιδράσεων στη διπλωμένη δομή, σχετίζεται με τη δημιουργία επιφανειών με συγκεκριμένες ιδιότητες, καταλλήλων για αναγνώριση συμπληρωματικών, καθώς διπλώνει -τοπικά- το τμήμα που περιέχει, υπό μορφή ενός στατιστικού, "ελαστικού" προτύπου, τα αμινοξέα που συμμετέχουν στις επιφάνειες αυτές.

Το ευχάριστο είναι ότι, αν οι επιτυχημένες περιπτώσεις σχεδιασμού, που περιγράφτηκαν στο Κεφ. Β.Ι, και κυρίως των DeGrado, Eisenberg και συνεργατών (δες Κεφ. Β.Ι για βιβλιογραφία), έχουν ο,τιδήποτε κοινό στη φιλοσοφία τους με αυτό που πραγματικά συνέβη εξελικτικά, αν δηλαδή δημιουργήθηκαν στην αρχή μικρότερα πεπτίδια, που ενώθηκαν σε μεγαλύτερα και βελτιστοποίησαν τη μεταξύ τους αλληλεπίδραση, τότε το σημαντικό ποσοστό των τυχαίων ολιγοπεπτιδίων, που είναι κατάλληλα για δεμάτι, γεφυρώνει την απόσταση μεταξύ των απλών δομικών λίθων, που δημιουργήθηκαν χημικά στην προβιωτική σούπα, και των μεγαλύτερων μορίων, όπως οι πρωτεΐνες: *οι επιμέρους συνιστώσες ήταν εφικτό να δημιουργη-θούν κατά τύχη*, και έμενε απλά να ενωθούν μεταξύ

τους σε μεγαλύτερες αλληλουχίες. Η οργάνωση σε ενότητες τοπικές κατά μήκος της αλληλουχίας, που είναι εύκολο να επιτευχθούν κατά τύχη, ενισχύει τη "σχεδιασιμότητα" των δομών σε επίπεδο αλληλουχίας, και την "επαναληψιμότητα" της σταθερότητας και λειτουργικότητας από οικογένεια σε οικογένεια. Η φιλοσοφία αυτή, που σχολιάζεται ήδη από το άρθρο ανασκόπησης του Dill [1990], αν και με κάποια ασάφεια ακόμη τότε, αποκτά όλο και περισσότερη υποστήριξη και έχει μόλις αρχίσει να γίνεται κατανοητή, μέσα από απλοποιημένα, θεωρητικά μοντέλα [Chan, 1998]. Δείχνει δε, ότι ο δρόμος του σχεδιασμού πρωτεϊνών, σαν αυτές που απαντούν στη φύση, ίσως να μην είναι τόσο δύσκολος· ενώ αντίθετα, αν υπάρχουν δομικά πρότυπα, που δεν έχει ανακαλύψει η εξέλιξη, τότε οι απαιτήσεις τους θα πρέπει να είναι τόσο συγκεκριμένες και πολύπλοκες, ώστε το δύσκολο θα είναι -μάλλον- να ξεστρατίσουμε από αυτά που μας διδάσκει η μηχανή της ζωής, προς άλλα άγνωστα αλλά και γοητευτικά μονοπάτια.

Στα μοντέλλα αυτά [Chan, 1998], φαίνεται ακόμη, ότι πρωτεΐνες, που έχουν τις περισσότερες αλληλεπιδράσεις οργανωμένες τοπικά, διπλώνουν ταχύτερα από άλλες με την ίδια ενέργεια σταθεροποίησης, αλλά με τις αλληλεπιδράσεις οργανωμένες μη-τοπικά. Αυτό δίνει μια μηχανιστική προοπτική στην τοπική οργάνωση μέρους της πληροφορίας, για τη δομή της κάθε πρωτεΐνης, όπως καταγράφηκε και για το δεμάτι. Επίσης ενισχύει την άποψη που διατυπώθηκε -κατ'αρχήν- στην Γενική Εισαγωγή, ότι η φυσική δομή δεν αποτελεί "αναγκαστικά" (τουλάχιστον όχι "απλά") το ολικό ενεργειακό ελάχιστο, αλλά -ίσως- ένα τοπικό ενεργειακό ελάχιστο εύκολα προσπελάσιμο από την κατάσταση αποδιάταξης. Αν σκεφτεί κανείς ότι οι αναλύσεις των Κεφ. Α.Ι-ΙΙ της παρούσας είχαν ολοκληρωθεί το 1992, σε συνδυασμό και με τα πειράματα άλλων ερευνητών [κυρίως Beauregard et al, 1995· Parker και Hefford, 1997· MacBeath et al, 1998], που έχουν αναφερθεί σε διάφορα σημεία ότι επιβεβαίωσαν και πειραματικά διάφορες πτυχές της παρούσας, αλλά πολύ αργότερα, και λάβει υπ'όψη του ότι τα θέματα αυτά τώρα γίνονται κατανοητά, καταλαβαίνει ότι πίνακες σαν αυτούς των Κεφ. Α.Ι-ΙΙ είναι μάλλον μόνο η αρχή…

Συνοψίζοντας

Η εργασία αυτή ξεκίνησε με στόχο τον εντοπισμό πιθανών προτιμήσεων των τοπολογικά διακριτών θέσεων, που μπορούν να οριστούν σε ένα τριτοταγές δομικό πρότυπο, για συγκεκριμένους αμινοξικούς τύπους. Αυτή η -από διαίσθηση ορμώμενη- ιδέα, που σχεδόν ασυναίσθητα έχει έντονο το "χρώμα" των μεμονωμένων (ίσως ανεξάρτητων;) θέσεων, έδωσε -από πολύ νωρίς- τη θέση της στην έννοια, όχι απλά της κατανομής συνολικά σε διάφορες θέσεις, αλλά της *ανακατανομής*, κυρίως στα πλαίσια των α-ελίκων που αποτελούν το δεμάτι, όπως αυτές περιγράφονται από την επανάληψη μιας βασικής επτάδας. Η σύσταση των επιμέρους θέσεων των ελίκων του δεματιού, σε αμινοξικούς τύπους, διαφέρει ξεκάθαρα (σε καλά επίπεδα στατιστικής σημαντικότητας, και πάντως πολύ περισσότερο από ότι των ελίκων συνολικά, που δεν διαφέρει ιδιαίτερα) από εκείνη των πρωτεϊνών που αποτελούνται από α-έλικες. Η στατιστική φύση των απαιτήσεων αυτών επιτρέπει, με μικρές μεταβολές των ποσοστών σε κάποιες θέσεις, να ρυθμίζονται επιμέρους θέματα της κάθε δομής, όπως απόσταση μεταξύ των α-ελίκων και ίσως και των μεταξύ τους γωνιών. Αντίθετα, σημαντικές μεταβολές σε προτιμήσεις για συγκεκριμένους αμινοξικούς τύπους σε συγκεκριμένες θέσεις, μπορεί να σημαίνει κάποια βασική αλλαγή στο πακετάρισμα (πχ. παράλληλο αντί για αντι-παράλληλο, όταν αυξάνουν τα β-διακλαδισμένα στη θέση εκείνη που τελικά θα είναι θέση a). Σε κάθε περίπτωση δεν αρκεί να υπάρχει μια συσσώρευση από κατάλοιπα που απλά ευνοούν το σχηματισμό α-έλικας, αλλά χρειάζεται και μια σωστή σειρά -υπό στατιστική, αλλά σε ορισμένες περιπτώσεις και μηχανιστική έννοια: ενώ στατιστικά οι περισσότερες αλλαγές δεν φέρνουν "την καταστροφή", μερικές φορές, αρκεί ένα λάθος κατάλοιπο για την πλήρη κατάρρευση του δομικού προτύπου ή τη σημαντική αλλαγή της δομής του. Το πλαίσιο που ορίζει η στατιστική φύση των πινάκων αντιστοιχεί στις *γενικεύσεις* που προκύπτουν για το δεμάτι σαν σύνολο, η δε χαλαρότητά τους, όπως επανειλημένα αναφέρθηκε, αντανακλά τη γενική φύση των περιορισμών που τίθενται λόγω πακεταρίσματος, σε αντίθεση με τους περιορισμούς που θέτει η τοπική διαμόρφωση της κύριας αλυσίδας, που συνήθως είναι αυστηροί και συγκεκριμένοι. Από την άλλη πλευρά, η ακρίβεια σε ένα μηχανιστικό επίπεδο αντιστοιχεί, όχι απλά στην *εξειδίκευση* για μια ολόκληρη οικογένεια, αλλά στην *εξατομίκευση* για μια συγκεκριμένη θέση a (ή d, ή ο,τιδήποτε) σε μια συγκεκριμένη αλυσίδα. Όσο χρήσιμη κι αν αποδεικνύεται η μεταφορά γνώσης μέσω αναλογίας ή επαγωγής, στο τέλος εκείνο που κρίνει την επιτυχία (και ενδεχόμενα τη ζωή ενός οργανισμού) είναι το αν θα διπλώσει και αν θα λειτουργήσει επαρκώς η συγκεκριμένη αλυσίδα!

Για τα συνδετικά τμήματα, αν και οι παραπάνω παρατηρήσεις επίσης ισχύουν, πιο έντονη είναι η σημασία του ταιριάσματος των διαφορετικών αναγκών σε μια -την κάθε φορά- λύση. Το δεμάτι δεν είναι απλά "τέσσερις α-έλικες με τρεις εύκαμπτους συνδέσμους ανάμεσα". Ειδικά οι μικρού μήκους συνδέσεις δείχνουν να ακολουθούν λίγες και πολύ συγκεκριμένες εναλλακτικές· σε αυτές τις περιπτώσεις, η τοπική διαμόρφωση (π.χ. μια θέση με θετική γωνία φ), μπορεί να παίζει καθοριστικό ρόλο στην επιλογή αμινοξικού τύπου (παρά το πακετάρισμα, ειδικά δοθέντος ότι πρόκειται για θέσεις στην επιφάνεια του μορίου). Οι ακραίες στροφές έχουν προτιμήσεις όχι μόνο για συγκεκριμένους αμινοξικούς τύπους, αλλά και για συγκεκριμένες θέσεις έναρξης (αντίστοιχα για το C-άκρο: λήξης)· και πάλι, οι θέσεις αυτές μπορεί να είναι υποχρεωτικές για περιπτώσεις σύντομων (άρα, με λίγες επιτρεπτές διαμορφώσεις) συνδέσεων, όμως οι ίδιες προτιμήσεις παρατηρούνται και με μεγαλύτερου μήκους συνδέσεις. Επιπλέον, συνήθως συμπίπτουν με την εγγύς προς κάποια γειτονική έλικα πλευρά. Ξεδιπλώνοντας αυτές τις χωροδιαταξικές ανάγκες από τον τρισδιάστατο χώρο, υπάρχουν σαφείς αντιστοιχίες στη μονοδιάστατη γραμμική αμινοξική αλληλουχία, όπου μάλιστα χρειάζεται προσοχή όταν συγκρίνονται συναφή πρότυπα, αφού οι ονομασίες αλλάζουν (πχ. η θέση a του δεματιού δεν είναι ίδια με τη θέση a της παράλληλης υπερέλικας).

Λεπτομερής πληροφορία του είδους, όπως πινακοποιείται στα παραπάνω κεφάλαια, μπορεί να διευκολύνει άμεσα όλο τα φάσμα των διαδικασιών πρωτεϊνικού σχεδιασμού, από απλές σημειακές και γενικότερα εντοπισμένες τροποποιήσεις, μέχρι το σχεδιασμό ολόκληρων πρωτεϊνικών μορίων από το μηδέν. Πρέπει να διευκρινιστεί ότι η βοήθεια αυτή εκφράζεται στο να δωθούν οι γενικότερες κατευθύνσεις, δηλαδή ποιες επιλογές είναι πιθανότερο να επιτύχουν. Το αρχικό γενικό σχέδιο θα πρέπει να εξατομικευτεί για την κάθε περίπτωση με συνδυαστικές διαδικασίες, που όμως καλούνται να διερευνήσουν σημαντικά (τάξεις μεγέθους) λιγότερους συνδυασμούς, αφού τα σημεία που αφήνονται για επιλογή με τέτοιες μεθόδους είναι λίγα. Έτσι το κυρίως βάρος μεταφέρεται στο να απαντηθούν συγκεκριμένα ερωτήματα, ή/και να σχεδιαστούν χρήσιμα προϊόντα.

Τελευταίο, αλλά ίσως το σημαντικότερο είναι ότι αναλύσεις του είδους, στον βαθμό που γενικεύονται, βοηθούν στην κατανόηση των σχέσεων ανάμεσα στην αλληλουχία και τη δομή των πρωτεϊνών, κάνοντας ξεκάθαρες έννοιες, όπου κάποιες αυτοματοποιημένες ρουτίνες έχουν -ίσως- ήδη σημειώσει επιτυχία, χωρίς όμως να μας έχουν πει ποτέ το πως.

“Έμεινα να αναρωτιέμαι: υπάρχει μοίρα ή περιπλανιόμαστε τυχαία, σαν την αύρα;
...ή μήπως και τα δυο;”

(Σε ελεύθερη απόδοση από την ταινία “Forrest Gump”)

Ανεξάρτητη Εργασία:

Συστοίχιση πρωτεϊνών βασισμένη σε εντοπισμένες ομοιότητες

Εισαγωγή

Κατά τη διάρκεια της κύριας εργασίας, που περιγράφηκε νωρίτερα, ανέκυψε επανειλημμένα ένα σοβαρό κενό στο χώρο των εργαλείων που χρησιμοποιούνται για την ανάλυση της δομής και λειτουργίας των πρωτεϊνών, και πιο συγκεκριμένα στα προγράμματα συστοίχισης/σύγκρισης αλληλουχιών πρωτεϊνών. Παρά τη μακρά ιστορία τους, οι σχετικοί αλγόριθμοι υποφέρουν ακόμη από διάφορα προβλήματα που οφείλονται στην ιδαίτερη φύση του καθενός. Οι διαδικασίες δυναμικού προγραμματισμού (dynamic programming)[31] εγγυώνται την εύρεση ενός βέλτιστου τρόπου συστοίχισης δύο αλληλουχιών, με βάση ένα δεδομένο πίνακα "ομοιοτήτων" μεταξύ των αμινοξικών τύπων (substitution table, similarity ή scoring matrix) και ένα "πρόστιμο" εισδοχών/απαλοιφών (ή "κενών", gap penalty), ακόμη και όταν οι αλληλουχίες αυτές δεν έχουν τίποτε κοινό μεταξύ τους. Η εκτίμηση, κατά πόσον αυτές πράγματι έχουν κάτι κοινό, γίνεται εκ των υστέρων και συνήθως με γνώμονα την πιθανότητα ένα τέτοιο επίπεδο ομοιότητας να έχει επιτευχθεί κατά τύχη. Αυτή η πιθανότητα όμως εξαρτάται όχι μόνο από το επίπεδο ομοιότητας, αλλά και το μήκος στο οποίο αυτή παρατηρείται. Έτσι, ένα μικρό αλλά έντονα όμοιο τμήμα θα μπορούσε να μην εντοπισθεί για χάρη ενός με περισσότερα κοινά χαρακτηριστικά αλλά σκορπισμένα σε μεγαλύτερη έκταση, ή απλά επειδή οι αλληλουχίες δεν παρουσιάζουν άλλα κοινά χαρακτηριστικά, οπότε είναι εύκολο να χαθεί στο θόρυβο της συστοίχισης των μη κοινών τμημάτων. Τα διαγώνια διαγράμματα (diagonal plot, dotplot), από την άλλη, αν και βολικά για εποπτικούς λόγους (ειδικά όπου εμπλέκονται εσωτερικές επαναλήψεις, μεταθέσεις δομικών ενοτήτων -domain shuffling- ή απλά υπάρχει μια μικρή περιοχή ομοιότητας μεταξύ δύο διαφορετικών κατά τ'άλλα αλληλουχιών), απαιτούν μια διαδικασία δοκιμής και σφάλματος μέχρι να επιλεγεί ο κατάλληλος συνδυασμός παραμέτρων που θα αναδείξει την επιθυμητή ομοιότητα.

[31] Η ορολογία εξηγείται όλη λεπτομερώς στο κυρίως κείμενο, όπως και η φιλοσοφία και οι αλγόριθμοι πίσω από τις διαδικασίες που αναφέρονται εδώ.

Παρουσιάζεται στη συνέχεια μια νέα προσέγγιση στο παλιό πρόβλημα της εύρεσης εντοπισμένων ομοιοτήτων (δηλαδή περιορισμένων σε μια μικρή ή μεγαλύτερη περιοχή κατά μήκος των δύο αλληλουχιών), βασισμένη στη βελτιστοποίηση τοπικά οποιουδήποτε μέτρου σημαντικότητας της συστοίχισης: ξεκινώντας από μικρά, στοιχειώδη και γενικά χωρίς σημασία κοινά τμήματα, δημιουργούνται όλο και μεγαλύτερα από μεταξύ τους συνενώσεις, μόνο όμως αν αυτά που προκύπτουν είναι με βάση κάποιο μέτρο καλύτερα. Το λογικό μοντέλο, πίσω από αυτή τη διαδικασία, είναι εκείνο του "διαγωνίου διαγράμματος". Μπορεί να χρησιμοποιηθεί με οποιοδήποτε μέτρο ομοιότητας σαν βάση, ενώ μπορεί να εφαρμοσθεί με τρόπους αποτελεσματικούς από πλευράς ταχύτητας και χώρου. Δεν χρειάζεται έξωθεν καθορισμό παραμέτρων, αφού η αυστηρότητα με την οποία γίνονται οι διάφορες επιλογές προσαρμόζεται αυτόματα στο επίπεδο της τοπικής ομοιότητας μεταξύ των δύο αλληλουχιών υπό σύγκριση. Το κυριότερο είναι ότι -ειδικά για συγκρίσεις μεταξύ πρωτεϊνών- ίσως να αντιστοιχεί με πιο φυσικό τρόπο στις διαδικασίες, που κρύβονται κάτω από τις ομοιότητες, που παρατηρούνται μεταξύ των αλληλουχιών τους, και με βάση τις οποίες (διαδικασίες) οι πρωτεΐνες αποκτούν την φυσική, λειτουργική, τρισδιάστατη δομή τους.

Μεθοδολογία

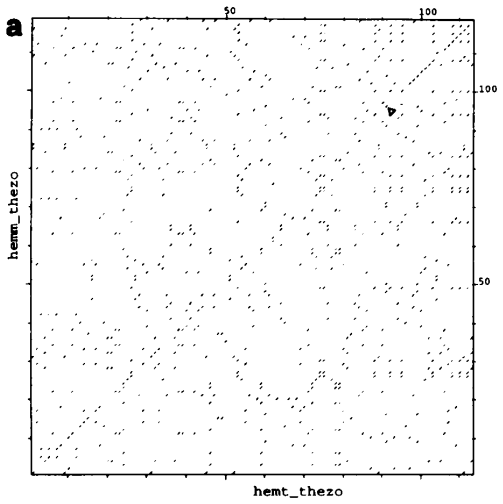
Χρήσιμοι ορισμοί: Η σχέση ανάμεσα σε δύο αλληλουχίες *A(i)* και *B(j)* που αποτελούνται από *m* και *n* αμινοξικά κατάλοιπα αντίστοιχα, μπορεί να περιγραφεί από ένα διδιάστατο πίνακα *R(i,j)*, μεγέθους *m*x*n*, τα στοιχεία του οποίου δείχνουν την ομοιότητα ανάμεσα στα κατάλοιπα στις θέσεις *i* της αλληλουχίας *A* και *j* της αλληλουχίας *B*. Προκειμένου για πρωτεΐνες, οι τιμές του πίνακα *R* προέρχονται από ένα πίνακα *S(x,y)*, μεγέθους 20x20, που δείχνει τη σχέση ομοιότητας ανάμεσα στους αμινοξικούς τύπους *x* και *y*. Συνήθως, ο πίνακας *S* έχει προκύψει εμπειρικά, με βάση πραγματικές πολλαπλές συστοιχίσεις [Dayhoff, 1978· Gonnet et al, 1992· Henikoff και Henikoff, 1992,1993]. Άλλα κριτήρια που μπορούν να χρησιμοποιηθούν περιλαμβάνουν ελάχιστο αριθμό αλλαγών βάσεων μεταξύ κωδικονίων, ή διαφορές μεταξύ των αμινοξικών τύπων σε σχέση με φυσικοχημικές και άλλες ιδιότητες. Προκειμένου για νουκλεοτιδικές αλληλουχίες, δίνεται μια βαθμολογία για τις ομοιότητες και μια για τις διαφορές (συνήθως 1 και 0 αντίστοιχα), αν και σε ορισμένες περιπτώσεις οι διαφορές βαθμολογούνται ελαφρά διαφορετικά. Στη συνέχεια, η σύγκριση παίρνει τη μορφή αναζήτησης, μέσα στον πίνακα *R*, διαγωνίων (ή τμημάτων τους) της μορφής (*R(i,j), R(i+1,j+1), ..., R(i+k,j+k)*) που να περιέχουν πολλά στοιχεία με βαθμολογία για υψηλή ομοιότητα (όχι κατ'ανάγκη όσο το δυνατό περισσότερα). Κατ'αρχήν εστιάζουμε στο πως η παρούσα προσέγγιση, βασισμένη σε αυτό το λογικό μοντέλλο, μπορεί να χρησιμοποιηθεί για συγκρίσεις μεταξύ πρωτεϊνών, ενώ στη συνέχεια δίνονται οι τροποποιήσεις που απαιτούνται για τη σύγκριση μεταξύ νουκλεοτιδικών αλληλουχιών. Επειδή, ανάλογα με το εκάστοτε πρόβλημα, μπορούν να χρησιμοποιηθούν διάφορες παραλλαγές, δίνεται μια γενική περιγραφή από την οποία μπορούν να προκύψουν διάφορες εφαρμογές, αντί για την αυστηρή αλλά "ξερή" περιγραφή ενός συγκεκριμένου αλγορίθμου. Μια σύντομη περιγραφή μιας συγκεκριμένης υλοποίησης, και περισσότερο με πνεύμα επεξηγήσεως, δίνεται αργότερα.

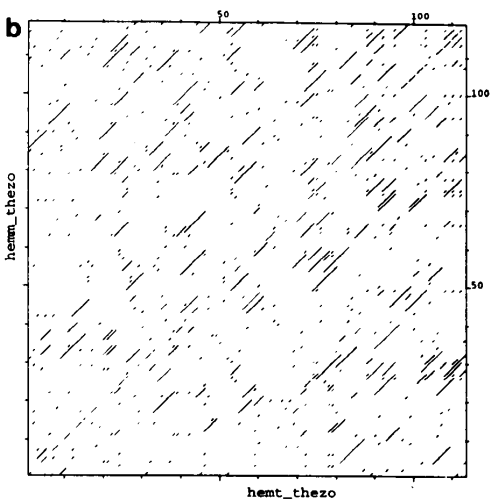
Γενική περιγραφή: Η προσέγγιση, που περιγράφεται εδώ, μπορεί να θεωρηθεί σαν μια διαδικασία δύο φάσεων. Στην πρώτη, καταχωρούνται, στον πίνακα *R*, μικρές, στοιχειώδεις, και γενικά μη-σημαντικές σημειακές ομοιότητες, συγκρίνοντας κάθε κατάλοιπο (ή μικρό τμήμα) της αλληλουχίας *A* με κάθε κατάλοιπο (ή μικρό τμήμα) της αλληλουχίας *B*. Στη δεύτερη φάση, αυτές οι σημειακές ομοιότητες συνενώνονται σε μεγαλύτερες υπο-συστοιχίσεις, που μπορούν να συνενωθούν επαναληπτικά σε όλο και μεγαλύτερες. Η διαδικασία μπορεί να σταματήσει όταν η σημαντικότητα της κάθε τοπικής συστοίχισης δεν βελτιώνεται παραπέρα (ή υπερβεί ένα όριο εμπιστοσύνης).

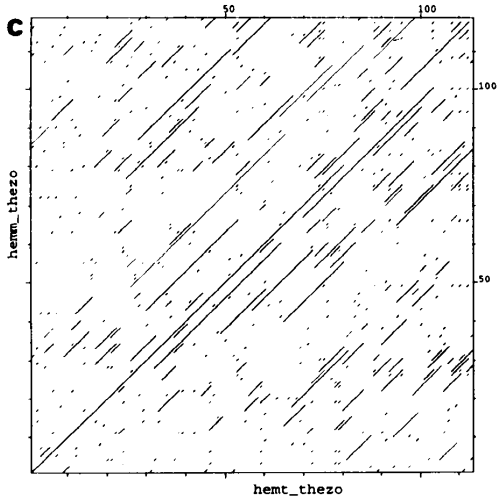
Στην πρώτη φάση (Εικ. 1a), και στην απλούστερη περίπτωση, μια αρχική σημειακή ομοιότητα μπορεί να ανιχνεύεται στο *R(i,j)* αν *A(i)=B(j)*. Στις πρωτεΐνες, η συχνότητα με την οποία συμβαίνει αυτό, σε διαγωνίους που δεν περιέχουν τμήματα συστοιχίσεων, είναι 5% (υποθέτοντας ίσες συχνότητες για τους είκοσι αμινοξικούς τύπους), ενώ μια πραγματική συστοίχιση δείχνει συνήθως τουλάχιστον 25% ταυτόσημα αμινοξικά

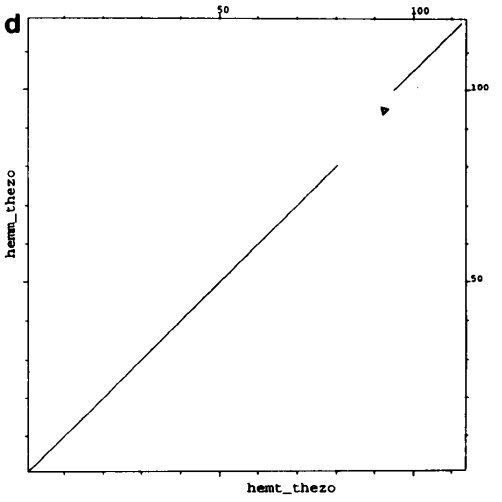
κατάλοιπα. Αν αντί για αμινοξική ταυτότητα, χρησιμοποιηθεί κάποιο μέτρο ομοιότητας (για παράδειγμα κάποιος από τους εμπειρικούς πίνακες ομοιότητας που αναφέρθηκαν παραπάνω), τότε μια αρχική σημειακή ομοιότητα μπορεί να ανιχνεύεται στο *R(i,j)* αν μια μικρή περιοχή γύρω από το *A(i)* παρουσιάζει συνολικά ομοιότητα με μια μικρή περιοχή γύρω από το *B(j)* περισσότερο από κάποιο κατώφλι. Τυπικά, 2-3 κατάλοιπα αρκούν, ενώ το κατώφλι δεν χρειάζεται να είναι μια αυθαίρετη εμπειρική τιμή, αλλά μπορεί να προσδιορίζεται από τον πίνακα ομοιότητας και το είδος της σύγκρισης.

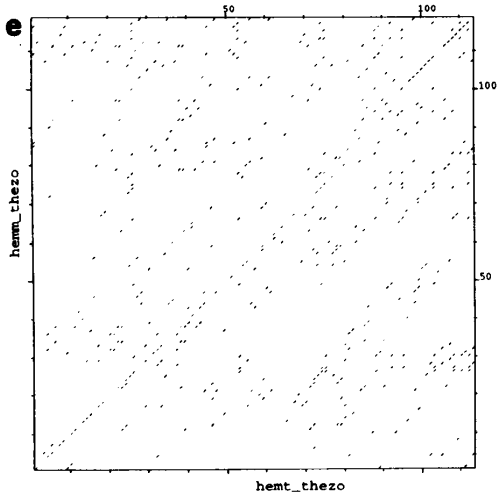
Στη δεύτερη φάση, και σε κάθε διαγώνιο, δύο γειτονικές (κατά μήκος της διαγωνίου) υπο-συστοιχίσεις μπορούν να συνενωθούν σε μια μεγαλύτερη, αν αυτή είναι καλύτερη σύμφωνα με οποιοδήποτε μέτρο από εκείνες από τις οποίες προέκυψε (Εικ. 1b). Αν το μέτρο είναι η στατιστική σημαντικότητα, τότε για το παραπάνω παράδειγμα που ξεκίνησε από τα ταυτόσημα κατάλοιπα, δύο γειτονικές αρχικές σημειακές ομοιότητες συνενώνονται σε μια μεγαλύτερη αν αυτή συμβαίνει με πιθανότητα λιγότερο από 5%, όταν συγκρίνονται άσχετες αλληλουχίες. Ο όρος γειτονικές αναφέρεται σε υπο-συστοιχίσεις που δεν χωρίζονται από άλλες κατά μήκος της διαγωνίου όπου ευρίσκονται· μπορεί να χωρίζονται από (τυπικά λίγα) αμινοξικά κατάλοιπα για τα οποία δεν ανιχνεύτηκε καμιά ομοιότητα, ή και να επικαλύπτονται, αν για τον εντοπισμό τους χρησιμοποιήθηκαν μικρά τμήματα αλληλουχιών. Ακριβώς εξαιτίας των πιθανών αποστάσεων/ επικαλύψεων, στη γενικότερη περίπτωση, η πιθανότητα να συμβεί η μεγαλύτερη υπο-συστοίχιση δεν είναι το γινόμενο των επιμέρους πιθανοτήτων. Σε κάθε περίπτωση, η τελική υπο-συστοίχιση εκτείνεται από την αρχή του πρώτου αρχικού τμήματος ομοιότητας μέχρι το τέλος του δεύτερου. Σε ακόλουθους κύκλους (Εικ. 1c,d), μπορεί κανείς να επιχειρήσει να συνενώσει τις νέες γειτονικές υπο-συστοιχίσεις σε μεγαλύτερες, πάνω στην ίδια βάση, και ξανά, μέχρι να επέλθει σύγκλιση (δηλαδή να μην μπορούν να γίνουν άλλες συνενώσεις) ή κάποιο επιθυμητό επίπεδο σημαντικότητας.











f

```
                96                  118
Hemm_Thezo ...kn VDYCKEWLVNHIKGTDFKYKGKL
                 :   : :::::::  :::::::
Hemt_Thezo ...gd VKWAKSWLVNHIKTIDFKYKGKI
                91                  113
```

Εικ. 1. *(Στην προηγούμενη σελίδα) Συστοίχιση ανάμεσα στην αιμερυθρίνη (κωδικός Hemt_Thezo) και την μυοαιμερυθρίνη (κωδικός Hemm_Thezo) από τον οργανισμό Themiste zostericola χρησιμοποιώντας το πρόγραμμα SPADEp (μια συστοίχιση που επελέγη τυχαία για επεξηγηματικούς λόγους). Δείχνεται η κατάσταση της διαδικασίας συστοίχισης σε επελεγμένους κύκλους, καθώς και το αποτέλεσμα ορισμένων "φίλτρων".*

α. Αρχική κατάσταση. Καθώς χρησιμοποιείται η ταυτότητα αμινοξικών τύπων σαν κριτήριο, μια κουκίδα τοποθετείται όπου ένα κατάλοιπο από την μία αλληλουχία είναι ταυτόσημο με κάποιο από την άλλη. Η κουκίδα που δείχνεται με το τριγωνικό βέλος επεξηγείται παρακάτω.

β. Τέλος του πρώτου κύκλου. Κάθε κουκίδα του προηγούμενου "καρέ" έχει ενωθεί με κάποια γειτονική, αν η συστοίχιση που προκύπτει είναι καλύτερη από την κάθε κουκίδα χωριστά.

γ. Τέλος του τελευταίου κύκλου. Κάθε τοπική συστοίχιση έχει επεκταθεί στον βέλτιστο συνδυασμό μήκους/ομοιότητας που μπορεί να επιτευχθεί με τη διαδικασία αυτή.

δ. Το ίδιο καρέ όπως στο γ, αλλά μετά την αφαίρεση υπο-συστοιχίσεων με σημαντικότητα μικρότερη του 4. Η θέση που που δείχνεται με το τριγωνικό βέλος επεξηγείται παρακάτω.

ε. Το ίδιο με το α, αλλά μετά την αφαίρεση όσων κουκίδων δεν γειτόνευαν με άλλες στην ίδια διαγώνιο, σε μια απόσταση επτά καταλοίπων σε καμιά από τις δύο κατευθύνσεις.

ζ. Συστοίχιση των καρβοξυτελικών ελίκων των δύο μορίων. Αμινοξικά κατάλοιπα που προηγούνται στην αλληλουχία δείχνονται με μικρά γράμματα. Η υπογραμμισμένη βαλίνη (που αντιστοιχεί στη κουκίδα που δείχνεται με το τριγωνικό βέλος στα α και δ) δεν ενώνεται με την λοιπή υπο-συστοίχιση, αφού 16 ταυτότητες σε 22 κατάλοιπα είναι λιγότερο σημαντικό από τις 15 στα 18. Ένα πιο ευαίσθητο σύστημα βαθμολόγησης, που θα ελάμβανε υπ'όψη και τις συντηρητικές μεταλλαγές στα ενδιάμεσα κατάλοιπα, ίσως να βοηθούσε στην επέκταση της υπο-συστοίχισης μέχρι τη βαλίνη αυτή.

Τεχνικά θέματα

Σημαντικότητα: Μέχρι εδώ αναφερθήκαμε στον όρο σημαντικότητα χωρίς άλλες διευκρινίσεις. Η στατιστική σημαντικότητα, όπως χρησιμοποιήθηκε στο παράδειγμα της προηγούμενης παραγράφου, θα μπορούσε να εξυπηρετήσει αυτό το σκοπό. Το πρόβλημα είναι ότι δεν υπάρχει πλήρης (αυστηρή) μαθηματική περιγραφή της στατιστικής σημαντικότητας γενικευμένων υπο-συστοιχίσεων με αυθαίρετους πίνακες ομοιοτήτων. Αφού επιλεγεί ο κατάλληλος πίνακας ομοιοτήτων, που μπορεί να είναι ένα σοβαρό πρόβλημα από μόνο του, η πιθανότητα να επιτευχθεί ένα επίπεδο ομοιότητας, σε μια υπο-συστοίχιση ενός δεδομένου μήκους, εξαρτάται από την κατανομή των είκοσι αμινοξικών τύπων κατά μήκος των υπό σύγκριση αλληλουχιών· αυτή η κατανομή δεν είναι ανεξάρτητη για επικαλυπτόμενα τμήματα. Τέλος, το γεγονός ότι οι αλληλουχίες των πρωτεϊνών δεν είναι τυχαιοποιημένες, περιπλέκει ακόμη περισσότερο τα πράγματα. Προσδιορισμός της στατιστικής σημαντικότητας μιας υπο-συστοίχισης με σάρωση μιας τράπεζας αλληλουχιών (όπου με τις περισσότερες αλληλουχίες λογικά δεν υπάρχει τίποτε κοινό), αίρει μέρος του προβλήματος. Όμως, οι τράπεζες αλληλουχιών, συνήθως δεν είναι τυχαιοποιημένες κατά μία ακόμη έννοια, αφού ορισμένες αλληλουχίες υπερ-αντιπροσωπεύονται. Αν λυθεί και αυτό το πρόβλημα με την επιλογή ενός μοναδικού συνόλου, όπου όλες οι αλληλουχίες αντιπροσωπεύονται το ίδιο, τότε παραμένει το πρόβλημα ότι ορισμένες υπο-αλληλουχίες που αντιστοιχούν σε κοινά δομικά χαρακτηριστικά (πχ. αμφιπαθικές α-έλικες) θα συναντώνται συχνότερα από άλλες.

Οι Sander και Schneider [1991] ανέπτυξαν ένα εμπειρικό κριτήριο για τη σημαντικότητα, εκφρασμένο σε όρους ταυτότητας αμινοξικών καταλοίπων επί τοις εκατό, που απαιτείται από μια υπο-συστοίχιση ενός δεδομένου μήκους, ώστε να υπονοείται δομική ομοιότητα. Προτείνουν ότι μια υπο-συστοίχιση μήκους L θα πρέπει, γι αυτό το σκοπό, να υπερβαίνει ένα κατώφλι $t(L)=290.15*L^{-0.562}$, για L μεταξύ 10 και 80. (Στην πραγματικότητα αυτός ο τύπος δίνει t(L)=100% για L=7 κατάλοιπα). Θα δούμε αργότερα, πως αυτή η πληροφορία αποδεικνύεται σημαντική σε σχέση με την προσέγγιση που περιγράφεται εδώ, αφού μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της σημαντικότητας ή μη σε δομικούς όρους. Δυστυχώς, αναζήτηση ενός αντίστοιχου συσχετισμού με ένα πιο ευαίσθητο σύστημα βαθμολογίας, που θα λάμβανε υπ'όψη του και συντηρητικές μεταλλαγές δεν έχει γίνει. Οι ίδιοι οι ερευνητές σχολιάζουν ότι, η υπο-συστοίχιση μπορεί να εντοπισθεί σε όρους ενός πιο ευαίσθητου συστήματος, και η σημαντικότητά της να εκτιμηθεί σε όρους ταυτότητας αμινοξικών καταλοίπων επί τοις εκατό, ένα μέτρο που δίνεται από όλους τους αλγορίθμους σύγκρισης αλληλουχιών.

Ορισμός των σημειακών (αρχικών) ομοιοτήτων: Αν και θεωρητικά οποιοδήποτε είδος σημειακών ομοιοτήτων θα μπορούσε να χρησιμοποιηθεί, η τελική εικόνα θα εξαρτάται σε μεγάλο βαθμό από τον ορισμό τους. Επειδή το βασικό χαρακτηριστικό της παρούσης προσέγγισης είναι η συνένωση γειτονικών υπο-συστοιχίσεων, αν δύο από αυτές χωρίζονται από μια τρίτη που δεν μπορεί να συνενωθεί με καμία από τις δύο, τότε δεν θα ενωθούν ούτε μεταξύ τους, ακόμη και αν το συνολικό αποτέλεσμα έχει μεγαλύτερη σημαντικότητα. Έτσι, για να επιλεγεί το καλύτερο δυνατό σημείο

εκκίνησης, η πιθανότητα με την οποία συμβαίνει το παραπάνω γεγονός θα πρέπει να εκτιμάται προσεκτικά, και στο πνεύμα της συγκεκριμένης εφαρμογής. Γενικά, οι αρχικές σημειακές ομοιότητες θα πρέπει να ορίζονται έτσι ώστε να μην συμβαίνουν συχνά σε διαγωνίους που δεν περιέχουν υπο-συστοιχίσεις, ενώ να δειγματοληπτούν επαρκώς και τις ασθενέστερες πιθανές υπο-συστοιχίσεις. Επιπλέον, θα πρέπει να αναζητάται το μικρότερο δυνατό μήκος αλληλουχίας (1-3 κατάλοιπα) με τον μεγαλύτερο δυνατό βαθμό ομοιότητας, αφού οι διαδοχικές συνενώσεις οδηγούν τις υπο-συστοιχίσεις σε όλο και μεγαλύτερα μήκη με όλο και μεγαλύτερη σημαντικότητα, αλλά και με όλο και χαμηλότερη μέση ομοιότητα ανά μήκος· οπότε, αν κάποιες αρχικές σημειακές υπο-συστοιχίσεις έχουν χαμηλό βαθμό ομοιότητας, θα είναι εύκολο να μπλοκάρουν την επέκταση, από τους πρώτους κιόλας κύκλους.

Αποτελεσματικότητα: Η προσέγγιση, που περιγράφεται στις προηγούμενες παραγράφους, δουλεύει, απαιτώντας χρόνο ανάλογο περίπου προς $\boldsymbol{O}(m \times n \times log_2 n)$, δίνοντας στο τέλος μια συλλογή όλων των χωρίς εισδοχές/απαλοιφές περιοχών ομοιότητας. Δίνονται στη συνέχεια διάφορες ιδέες που μπορούν να αυξήσουν την αποτελεσματικότητα σε θέματα ταχύτητας και απαιτούμενου χώρου σε διάφορες υλοποιήσεις.

Για παράδειγμα, για ορισμένα είδη (αρχικών) σημειακών ομοιοτήτων, μπορεί να είναι εκμεταλλεύσιμοι πίνακες κατακερματισμού, που βοηθούν να γίνεται ο εντοπισμός τους σε χρόνο ανάλογο προς $\boldsymbol{O}(m+n)$ και ταυτόχρονα υποδεικνύουν τις διαγωνίους εκείνες στις οποίες πρέπει να αναζητηθούν οι πιθανές συστοιχίσεις. (Αυτό, παρεμπιπτόντως, είναι το "μυστικό" της ταχύτητας του ευρύτατα χρησιμοποιημένου, για αναζητήσεις ομόλογων αλληλουχιών σε τράπεζες πληροφοριών, προγράμματος FASTA [Wilbur και Lipman, 1983· Lipman και Pearson, 1985· Pearson και Lipman, 1988].)

Διάφορα φίλτρα μπορούν να εφαρμοσθούν στο τέλος της αρχικής φάσης, ώστε να απορρίπτονται οι αρχικές σημειακές ομοιότητες που κρίνονται με βάση κάποιο κριτήριο ότι δεν είναι πιθανό να συμμετέχουν σε σημαντικές υπο-συστοιχίσεις. Σαν μια χονδρική περίπτωση, στο παραπάνω παράδειγμα, όπου χρησιμοποιήθηκε η ταυτότητα αμινοξικών τύπων, θα μπορούσε να απορριφθεί, χωρίς συνέπειες για την ευαισθησία της σύγκρισης, οποιαδήποτε αρχική ομοιότητα δεν γειτονεύει με άλλη σε μήκος 7 καταλοίπων μπρος και πίσω, αφού μια περιοχή 15 καταλοίπων με ένα μόνο κοινό, δεν μπορεί να συμμετέχει σε οποιαδήποτε σημαντική συστοίχιση.

Επίσης, ορισμένα είδη σημειακών ομοιοτήτων, όταν είναι διαδοχικά στην αλληλουχία, ή επικαλυπτόμενα, η συνένωσή τους δίνει εξ ορισμού υπο-συστοιχίσεις μεγαλύτερης σημαντικότητας. Αυτές οι περιπτώσεις θα μπορούσαν να λαμβάνονται υπ'όψη χωριστά σε ένα προ-στάδιο, πριν από τη δεύτερη φάση, και να οδηγούνται σε συνένωση χωρίς υπολογισμούς σημαντικότητας που μπορούν να είναι χρονοβόροι.

Στη γενικώτερη περίπτωση πάντως, επικαλυπτόμενες γειτονικές υπο-συστοιχίσεις δεν έχουν αθροιστικά καλύτερη σημαντικότητα. Αυτό συμβαίνει όταν η μία από τις δύο έχει προκύψει από

συνένωση μικρότερων τμημάτων με υψηλό βαθμό ομοιότητας, και η άλλη από αντίστοιχα χαμηλού βαθμού ομοιότητας. Αυτό μπορεί να αποτελέσει ιδιαίτερη βοήθεια στον εντοπισμό περιοχών μικρού μήκους με υψηλό βαθμό ομοιότητας, ή περιοχών με απροσδόκητο βαθμό διαφοράς σε παρόμοιες κατά τα άλλα αλληλουχίες. Όμως, αν τέτοιου είδους πληροφορία δεν απαιτείται, τότε, από τη στιγμή που διάφορες διαδοχικές υπο-συστοιχίσεις έχουν όλες υπερβεί ένα όριο σημαντικότητας, μπορούν να συνενώνονται σε μια συνολική συστοίχιση χωρίς άλλους υπολογισμούς σημαντικότητας.

Τέλος, μπορεί κανείς να αλλάξει από τη διαδικασία "εκτίμησης-και-συνένωσης" σε οποιοδήποτε επιθυμητό σχήμα στο τέλος οποιουδήποτε κύκλου. Ένα προφανές τέτοιο σημείο απορρέει από το γεγονός ότι οι τελικές (υπο-)συστοιχίσεις κινούνται σε όρια αρχικών σημειακών ομοιοτήτων· δηλαδή η καθεμία τους ξεκινά στην αρχή μιας τέτοιας και τελειώνει στο τέλος μιας άλλης. Ίσως να θέλει κανείς, για παράδειγμα, να ερευνήσει αν η συστοίχιση θα μπορούσε να επεκταθεί κι άλλο κατά μήκος της διαγωνίου, αν και με ένα ευαίσθητο ορισμό των αρχικών σημειακών ομοιοτήτων αυτό θα ήταν σπάνιο. Ή θα μπορούσε κανείς να επιχειρήσει να ενώσει τις τελικές υπο-συστοιχίσεις που χωρίζονται από εισδοχές/απαλοιφές χρησιμοποιώντας μια έκδοση του αλγορίθμου των Needdleman και Wunsch [1970], "περιορισμένη" ανάμεσα στα άκρα των χωρίς κενά υπο-συστοιχίσεων. Θα πρέπει όμως να θυμάται κανείς ότι οι εισδοχές/απαλοιφές αντιστοιχούν συνήθως σε τμήματα δομής συνδετικά μεταξύ των στοιχείων ομαλής β'-ταγούς δομής, με διαφορετικό μήκος, και όχι κατ'ανάγκη δομικά ισοδύναμα.

Απαιτούμενες τροποποιήσεις για τη σύγκριση νουκλεοτιδικών αλληλουχιών: Οι ίδιες αρχές που διέπουν αυτή την προσέγγιση για πρωτεΐνες, ισχύουν και για τις συγκρίσεις μεταξύ νουκλεοτιδικών αλληλουχιών. Μια βασική διαφορά είναι ότι πρέπει να χρησιμοποιηθούν κάπως μεγαλύτερου μεγέθους αρχικές σημειακές ομοιότητες σαν σημείο εκκίνησης. Μια καλή επιλογή θα μπορούσε να είναι δύο ταυτόσημα νουκλεοτίδια ακολουθούμενα από άλλα δύο, αμέσως ή με ένα αταίριαστο νουκλεοτίδιο ενδιάμεσα. Αυτό θα βοηθούσε ιδιαίτερα στη συστοίχιση κωδικών περιοχών, όπου η τρίτη βάση είναι γενικά πιο μεταλλάξιμη. Αυτή η επιλογή, σε διαγωνίους που δεν περιέχουν υπο-συστοιχίσεις, δίνει μια σημειακή ομοιότητα κάθε 147 νουκλεοτίδια· οι περισσότερες από αυτές απορρίπτονται από το φιλτράρισμα στο τέλος της πρώτης φάσης. Ένα δεύτερο πρόβλημα, αφορά το γεγονός ότι εδώ δεν υπάρχει μέτρο σημαντικότητας αντίστοιχο της δομικής ομοιότητας των πρωτεϊνών· το μόνο διαθέσιμο μέτρο είναι η στατιστική σημαντικότητα με όλα τα προβλήματα που συνεπάγεται. Βαριά βιολογική γνώση πρέπει να χρησιμοποιείται στη συνέχεια για να γίνει διάκριση τι είναι πραγματικό και τι απλά στατιστικά σημαντικό.

Υλοποίηση

Οι παραπάνω αρχές, υλοποιήθηκαν με τη μορφή προγραμμάτων για ηλεκτρονικό υπολογιστή. Τα προγράμματα SPADEp και SPADEn (Search Procedure for All Diagonal Elements in proteins/nucleic acids αντίστοιχα), προορίζονται για συγκρίσεις μεταξύ δύο αλληλουχιών, είναι δε σχεδιασμένα να έχουν ελάχιστες απαιτήσεις (πχ. αρκούν ένα τερματικό κειμένου και η δυνατότητα να εκτυπώνει κανείς αρχεία Postscript· επιπλέον, αν και όλες οι τοπικές υπο-συστοιχίσεις διατηρούνται διαρκώς στη μνήμη RAM, οι απαιτήσεις σε μνήμη είναι της τάξης του 1MB για συγκρίσεις μεταξύ αλληλουχιών μήκους μέχρι 3000 περίπου βάσεων). Δέχονται αλληλουχίες διαφόρων τυποποιήσεων (PIR, GCG, STADEN, ακόμη και είσοδο από πληκτρολόγιο), ενώ παράγουν αποτελέσματα τόσο σε μορφή κειμένου, όσο και γραφικών. Άλλες λεπτομέρειες της λειτουργίας περιλαμβάνουν οργάνωση της ηλεκτρονικής βοήθειας έτσι ώστε να έχει κανείς ανά πάσα στιγμή τις πληροφορίες που χρειάζεται για αυτό που συμβαίνει εκείνη τη στιγμή, και μόνο. Επιπλέον, επειδή -από τη φύση της προσέγγισης- όλες οι τοπικές υπο-συστοιχίσεις διατηρούνται διαρκώς στη μνήμη, ενθαρρύνεται ο πειραματισμός με διάφορα φίλτρα, ενώ αποτελέσματα και στις δύο μορφές (κείμενο και γραφικά) μπορούν να παραχθούν οποιαδήποτε στιγμή (όλα τα κομμάτια της Εικόνας 1 δημιουργήθηκαν σε μια μόνο εκτέλεση του προγράμματος).

Στο πιο επιστημονικό μέρος, και για τις εκδόσεις αυτές, οι αρχικές σημειακές υπο-συστοιχίσεις, ορίζονται με βάση τα ταυτόσημα κατάλοιπα για τις πρωτεϊνικές, και με βάση ταυτόσημα τετρανουκλεοτίδια για τις νουκλεοτιδικές αλληλουχίες. Αυτή η διαισθητική επιλογή αποδείχτηκε να συμπεριφέρεται αρκετά καλά, στην πρώτη αυτή προσέγγιση. Για τον ταχύτερο εντοπισμό τους, χρησιμοποιείται ένας πίνακας κατακερματισμού και άμεσης αναζήτησης (hash-code-and-lookup table), ενώ ένας απλός μηχανισμός ατζέντας εξασφαλίζει ότι δεν ελέγχονται διαγώνιοι στις οποίες έχει ήδη επέλθει σύγκλιση.

Η στατιστική σημαντικότητα εκτιμάται με βάση τον τύπο των Altschul και Erickson [1986], χρησιμοποιώντας όμως ομοιότητα αντί για διαφορές. Πιο συγκεκριμένα, το μέτρο που χρησιμοποιείται είναι ο αρνητικός λογάριθμος με βάση το 20 για τις πρωτεΐνες (ή το 4 για τα ναουκλεϊκά) της πιθανότητας να επιτύχει κανείς κατά τύχη k ταυτόσημα κατάλοιπα (βάσεις) σε μήκος n, υποθέτοντας ίδιες συχνότητες καταλοίπων (βάσεων) και δυωνυμική κατανομή:

$$P\text{-value} = -\log_{1/p}\Sigma p^{i}(1-p)^{n-i}\, n!/i!(n-i)!\ ,\ i=k\ldots n \qquad (1)$$

Αυτός ο τύπος αγνοεί το γεγονός ότι η πιθανότητα αυτή δεν είναι ανεξάρτητη για επικαλυπτόμενες υπο-αλληλουχίες. Έτσι, αν για παράδειγμα μια υπο-συστοίχιση μεταξύ πρωτεϊνών έχει μια τιμή P_{20}=7 σημαίνει ότι είναι τόσο δύσκολο να επιτευχθεί κατά τύχη (δηλαδή μεταξύ μη ομολόγων αλληλουχιών) όσο το να βρεθούν 7 συνεχόμενα κοινά κατάλοιπα. Σημειωτέον, ότι οι συνδυασμοί μηκών και επί τοις 100 ταυτότητας που προτείνονται από το κριτήριο των Sander και Schneider [1991], αντιστοιχούν συστηματικά σε τιμές P_{20} γύρω στο 6.7 . Αυτό εξηγείται ως εξής: επειδή η τυπική απόκλιση της δυωνυμικής κατανομής είναι $(n\cdot p\cdot(1-p))^{0.5}$, η μετατροπή στη συνήθως

χρησιμοποιούμενη βαθμολογία z (δηλαδή τον αριθμό των τυπικών αποκλίσεων μακριά από τον μέσο όρο μιας κατανομής) είναι άμεση και είναι $(k-n\cdot p-0.5)/ (n\cdot p\cdot(1-p))^{0.5}$. Για πρωτεΐνες, αυτό προσεγγίζεται από $k/(n\cdot p)^{0.5}$, αφού $n\cdot p<<k$ και από κάποιο μήκος και πάνω $0.5<<k$ και $1-p=0.95\sim1$. Οπότε, επειδή το ποσοστό ταυτότητας είναι k/n, ο τύπος των Sander και Schneider [1991] μπορεί να διατυπωθεί $k/n^{0.438}>2.9$. Διαιρώντας και τα δύο μέλη με $p^{0.5}=0.223$, γίνεται $k/(n^{0.438}\cdot p^{0.5})>13$, το αριστερό μέρος της οποίας προσεγγίζει καλά τη διατύπωση σε όρους βαθμολογίας z του τύπου [1]. Πρέπει να σημειωθεί ότι (α) αυτή είναι μια ποιοτική εξήγηση του φαινομένου και όχι μια μαθηματική απόδειξη (β) η κανονική κατανομή αποκλίνει στην ουρά μακριά από το μέσο όρο από τη δυωνιμική κατανομή, που σημαίνει ότι η βαθμολογία z δεν μεταφράζεται άμεσα αντίστροφα σε πιθανότητες μέσω της κανονικής κατανομής. Από την άλλη πλευρά όμως, η βαθμολογία z είναι μονότονη επί της πιθανότητας (όσο ψηλότερη η βαθμολογία z, τόσο χαμηλότερη η πιθανότητα), που σημαίνει ότι προσεγγίσεις υπολογιστικά φθηνότερες από τον υπολογισμό της δυωνιμικής μπορεί να είναι χρήσιμες στη φάση της σύγκλισης. Τέλος, οι συστοιχίσεις από τις οποίες προέκυψε το κριτήριο των Sander και Schneider [1991] περιελάμβαναν κενά, που αποτελεί μια ένδειξη ότι ίσως οι παραπάνω επιλογές να γενικεύονται και σε αυτές τις περιπτώσεις.

Επαναπροσδιορισμός των αρχικών σημειακών συστοιχίσεων

Στην περίπτωση των πρωτεϊνών, η παραπάνω συζήτηση, γύρω από την αντιστοιχία μεταξύ της στατιστικής σημαντικότητας και της δομικής ομοιότητας, προσφέρει έναν ακόμη τρόπο προσδιορισμού των αρχικών σημειακών συστοιχίσεων. Αυτές μπορούν να οριστούν έτσι ώστε, μετά από 1-2 κύκλους συνενώσεων, να οδηγούν σε υπο-συστοίχισεις που να αντιστοιχούν με μεγάλη πιθανότητα σε δομικά παρόμοια τμήματα. Έχει δειχθεί [Argos, 1987β] ότι πενταπεπτίδια με αλληλουχία ταυτόσημη ή με μια μόνο διαφορά υιοθετούν διαφορετική δομή σε διαφορετικά περιβάλλοντα, αν και το Ό των περιπτώσεων έδειχνε καθαρή προτίμηση για μια διαμόρφωση. Αν και ταυτόσημα τετρα- ή πενταπεπτίδια είναι μια πολύ αυστηρή επιλογή για τον ζητούμενο σκοπό, γενικά αυτή η πληροφορία, σε συνδυασμό με το κριτήριο των Sander και Schneider [1991], δίνουν μια αίσθηση του μεγέθους και της απαιτούμενης ομοιότητας για τις αρχικές σημειακές συστοιχίσεις. Στην περίπτωση των νουκλεοτιδικών αλληλουχιών, ένα αντίστοιχο μέτρο θα μπορούσε να δωθεί από μια βιβλιοθήκη ενδιαφερόντων προτύπων (τα συνήθως αποκαλούμενα κουτιά -boxes)· αυτή η βιβλιοθήκη θα μπορούσε να περιέχει όχι μόνο τις αλληλουχίες των προτύπων αυτών, αλλά και τις σχετικές τους συχνότητες, καθώς και ενδεχόμενες σχέσεις με άλλα πρότυπα (συχνά κάποια από αυτά προηγούνται ή έπονται άλλων).

Είναι φανερό, λοιπόν, ότι η ιδέα των γειτονικών μικρότερων υπο-συστοιχίσεων που ενώνονται σε όλο και μεγαλύτερες, σε συνδυασμό και με την παραπάνω άποψη για τις αρχικές σημειακές συστοιχίσεις (σε σχέση με την δομική ομοιότητα), αντιστοιχούν με πιο φυσικό τρόπο στους μηχανισμούς που κρύβονται πίσω από τις ομοιότητες που παρατηρούνται ανάμεσα στις βιολογικές αλληλουχίες: β'-ταγή στοιχεία, που οργανώνονται σε δομικά πρότυπα, που οργανώνονται σε ανώτερης τάξης δομές.

Συζήτηση

Διαφορές από άλλες προσεγγίσεις

Διαγώνια διαγράμματα: Η παρούσα προσέγγιση θυμίζει έντονα την κατηγορία αυτή, ειδικά αν για τις αρχικές σημειακές συστοιχίσεις χρησιμοποιηθούν μικρά τμήματα από τις υπό σύγκριση αλληλουχίες. Στις προσεγγίσεις αυτής της κατηγορίας, χρησιμοποιούνται "παράθυρα" σταθερού μεγέθους (fixed-size windows): η τιμή *R(i,j)* στον πίνακα της σύγκρισης, είναι η βαθμολογία που προκύπτει από την σύγκριση ενός μικρού τμήματος (παραθύρου) γύρω από το κατάλοιπο *i* της μιας αλληλουχίας, με ένα μικρό τμήμα ίσου μεγέθους γύρω από το κατάλοιπο *j* της άλλης αλληλουχίας. Ένα σημείο τοποθετείται στο διάγραμμα στο σημείο *(i,j)* αν η βαθμολογία αυτή ξεπεράσει ένα ελάχιστο απαιτούμενο όριο. Η εικόνα που προκύπτει είναι το λεγόμενο "διαγώνιο διάγραμμα" (diagonal plot, ή dotplot), όπου διαδοχικές "κουκίδες" κατά μήκος μιας διαγωνίου δείχνουν υπο-συστοίχιση. Όμως η απόφαση για την τοποθέτηση της κάθε κουκίδας έχει παρθεί ανεξάρτητα.

Τα διαγώνια διαγράμματα είναι πολύ καλά από την εποπτική πλευρά, αλλά συνήθως πρέπει να δοκιμαστεί ένας αριθμός συνδυασμών μεγεθών παραθύρου και βαθμολογιών αποδοχής: μικρά μεγέθη παραθύρου με αυστηρή απαιτούμενη βαθμολογία, δίνουν μια καθαρή εικόνα, υψηλής ευκρίνειας, αλλά είναι κατάλληλα για πρωτεΐνες με μεγάλη ομοιότητα· αντίθετα, μεγαλύτερα μεγέθη παραθύρου με πιο επιεικείς βαθμολογίες αποδοχής πρέπει να χρησιμοποιηθούν για να εντοπισθούν κοινά χαρακτηριστικά μεταξύ πρωτεΐνών με μικρότερο βαθμό ομοιότητας. Όταν δύο αλληλουχίες έχουν μεγάλη ομοιότητα σε μια μικρή περιοχή, και χαμηλότερη στο υπόλοιπο μήκος τους, συχνά αυτό είναι δύσκολο να δειχτεί με ένα μόνο συνδυασμό παραθύρου/βαθμολογίας. Είναι δυνατό [Argos, 1987α] να δοκιμάζονται σε κάθε σημείο του διαγράμματος συστηματικά όλοι οι συνδυασμοί παραθύρου/βαθμολογίας, και να αποδίδεται στο σημείο αυτό η υψηλότερη βαθμολογία z που επιτεύχθηκε, ανεξάρτητα από ποιον συνδυασμό έγινε αυτό. Χρησιμοποιώντας διαφορετικά σύμβολα για κάθε επίπεδο z, τα σημεία όπου είναι πιο αξιόπιστη η ομοιότητα μπορούν να εντοπιστούν με μια ματιά. Οι Altschul και Erickson [1986] περιγράφουν μια διαδικασία παρόμοια με την παρούσα (τον αλγόριθμο DD), όπου όλα τα διαδοχικά σημεία ταυτότητας μεταξύ των δύο αλληλουχιών, που βρίσκονται στην ίδια διαγώνιο, ενώνονται σε τμήματα που εξυπηρετούν ένα ρόλο αντίστοιχο με τις σημειακές συστοιχίσεις από τις οποίες αρχίζει και η παρούσα, μόνο που γενικά είναι πιο μεγάλα. Σε μια δεύτερη φάση, δοκιμάζονται συστηματικά όλες οι πιθανες $r(r+1)/2$ υπο-συστοιχίσεις που ξεκινούν από την αρχή ενός τέτοιου τμήματος και τελειώνουν στο τέλος ενός άλλου, στην ίδια όμως διαγώνιο, και στο τέλος απεικονίζονται όσες υπο-συστοιχίσεις ξεπερνούν ένα όριο σημαντικότητας. Το βασικό πρόβλημα των αλγορίθμων της οικογένειας των διαγωνίων διαγραμμάτων, είναι η ταχύτητα: ο χρόνος, που απαιτείται για τη σύγκριση δύο αλληλουχιών, είναι ανάλογος προς mxnx(μέγεθος παραθύρου)x(αριθμός συνδυασμών παραμέτρων που δοκιμάζονται), περιλαμβάνοντας πάντα τον επιπλέον χρόνο που χρειάζεται για τον προσδιορισμό του επιπέδου του θορύβου, για κάθε συνδυασμό, αν χρησιμοποιηθεί βαθμολογία z. Αυτό το πρόβλημα καθορίζει και το πλαίσιο στο οποίο συνήθως χρησιμοποιείται ένας τέτοιος αλγόριθμος, δηλαδή συγκρίσεις μεταξύ δύο

αλληλουχιών μόνο, ειδικά όπου υπάρχουν εσωτερικές επαναλήψεις ή μεταθέσεις δομικών ενοτήτων, ή για να απεικονισθεί η ομοιότητα σε ένα μικρό τμήμα μεταξύ δύο διαφορετικών κατά τ'άλλα αλληλουχιών. Η παρούσα προσέγγιση χρησιμοποιεί τις έτοιμες, από προηγούμενους κύκλους, υπο-συστοιχίσεις, και τις ενώνει σε όλο και μεγαλύτερες, ελέγχοντας ταυτόχρονα τη σημαντικότητά τους, σαν να κινείται σταδιακά (και μέσα σε λίγους κύκλους στην πράξη) προς τον βέλτιστο συνδυασμό παραθύρου/βαθμού ομοιότητας και μάλιστα χωριστά για το κάθε σημείο του διαγράμματος.

Δυναμικός προγραμματισμός: Όταν χρησιμοποιούνται πίνακες ομοιότητας $S(x,y)$ τέτοιοι ώστε κάποια λίγα στοιχεία του πίνακα να είναι θετικά (πχ. $S(x,y)>0$ όπου υπάρχει ταυτότητα αμινοξικών τύπων, δηλαδή $x=y$) ενώ ο σταθμισμένος μέσος όρος του πίνακα είναι αρνητικός:

$$\Sigma\Sigma p_x p_y S(x,y) < 0$$

(όπου p είναι οι συχνότητες με τις οποίες απαντούν οι διάφοροι αμινοξικοί τύποι, και τα αθροίσματα υπολογίζονται για όλα τα x,y), τότε το άθροισμα κατά μήκος οποιασδήποτε διαγωνίου του πίνακα R θα γίνεται σταδιακά όλο και πιο αρνητικό, εκτός αν η διαγώνιος περιέχει κάποια υποσυτοίχιση. Αυτή η παρατήρηση έχει οδηγήσει στη χρήση του δυναμικού προγραμματισμού σαν μια μέθοδο για την ανίχνευση τοπικών υπο-συστοιχίσεων, όπως υλοποιήθηκε από τους Smith και Waterman [1981] σε μια "τοπική" παραλλαγή του αρχικού αλγορίθμου των Needleman και Wunsch [1970]. Σε αυτή την προσέγγιση, οι συστοιχίσεις μπορούν να περιλαμβάνουν κενά, που στο διαγώνιο διάγραμμα αντιστοιχούν σε πέρασμα από μια διαγώνιο σε κάποια άλλη, οπότε και αφαιρείται ένα προκαθορισμένο "πρόστιμο" από τη συνολική βαθμολογία. Η παρούσα προσέγγιση, όπου σε διαδοχικούς κύκλους όλο και μεγαλύτερα τμήματα χρησιμοποιούνται, διαφέρει σαφώς από την προσέγγιση αυτή όπου σε διαδοχικούς κύκλους η συστοίχιση προχωρά κατά ένα κατάλοιπο την κάθε φορά.

Οι Altschul και συνεργάτες περιέγραψαν λεπτομερώς τις ιδιότητες των πινάκων ομοιότητας [Karlin και Altschul, 1990· Altschul, 1991]. Έδειξαν μεταξύ άλλων ότι κάθε πίνακας μπορεί να μετατραπεί σε μια μορφή που να ικανοποιεί τα παραπάνω δύο κριτήρια (δηλαδή μερικά τουλάχιστον θετικά στοιχεία και αρνητικό αναμενόμενο μέσο όρο). Διαδικασίες που εκμεταλλεύονται αυτά τα κριτήρια, πέρα από το πρόστιμο για το άνοιγμα ενός κενού στη στοίχιση (που είναι ένα κεφάλαιο από μόνο του) έχουν το πρόβλημα ότι βασίζονται στην αθροιστική βαθμολογία για τον εντοπισμό της καλύτερης στοίχισης, ενώ για την εκτίμηση της σημαντικότητάς της βασίζονται σε άλλα κριτήρια, όπως η στατιστική σημαντικότητα. Αυτή όμως εξαρτάται όχι μόνο από την αθροιστική βαθμολογία, αλλά και το μήκος στο οποίο αυτή επιτυγχάνεται. Έτσι, ένα μικρού μήκους κοινό τμήμα με έντονη ομοιότητα μεταξύ των δύο αλληλουχιών μπορεί να μην εντοπισθεί για χάρη ενός με υψηλότερη βαθμολογία αλλά σε μεγαλύτερο μήκος, δηλαδή με χαμηλότερη μέση ομοιότητα και πιθανά λιγότερο σημαντικό. Μάλιστα, ακόμη και ο προσδιορισμός της σημαντικότητας στοιχίσεων που περιέχουν κενά είναι τόσο δύσκολος που συνήθως γίνεται με προσωμοίωση: δημιουργούνται τυχαιοποιημένες

αλληλουχίες με τη σύσταση σε αμινοξικά κατάλοιπα της μιας από τις δύο, οι οποίες συστοιχίζονται με την άλλη, και το αποτέλεσμα της κανονικής συστοίχισης συγκρίνεται με το μέσο όρο και την τυπική απόκλιση της κατανομής τους. Μόνο πρόσφατα [Waterman και Vingron, 1994] έχει προταθεί μια ημι-εμπειρική προσέγγιση στο πρόβλημα, που απαιτεί μόνο μια προσωμοίωση για ένα δεδομένο πίνακα ομοιοτήτων και ένα πρόστιμο για το άνοιγμα κενών αντί για μια προσωμοίωση για κάθε συστοίχιση, όπως συνήθως γίνεται. Στην παρούσα προσέγγιση, όπου η συνένωση των υπο-συστοιχίσεων σε μεγαλύτερες οδηγείται με βάση τον έλεγχο της σημαντικότητας σε κάθε κύκλο, η τελική έκταση των υπο-συστοιχίσεων είναι απλά θέμα σύγκλισης, και χρησιμοποιείται το ίδιο μέτρο και για τον εντοπισμό και για την εκτίμηση των υπο-συστοιχίσεων, αντί για δύο. Το μέτρο που χρησιμοποιείται είναι απλοϊκό, αλλά έχει πειραματική υποστήριξη, μπορεί να υλοποιηθεί σε υπολογιστικά φθηνές διαδικασίες, δεν απαιτεί προσωμοιώσεις, και φαίνεται χρήσιμο ακόμη και στις περιπτώσεις συστοιχίσεων που περιλαμβάνουν κενά.

Άλλα προβλήματα των προγραμμάτων της οικογένειας του δυναμικού προγραμματισμού περιλαμβάνουν την αδυναμία να ληφθούν υπ'όψη εσωτερικές επαναλήψεις, ή μεταθέσεις δομικών ενοτήτων, ενώ το πρόβλημα του απαιτούμενου χρόνου και χώρου έχει επίσης τεθεί από ορισμένους. Και πάλι είναι το πλαίσιο στο οποίο χρησιμοποιούνται τα προγράμματα αυτά που καθιστά τα προβλήματά τους "όχι ιδιαίτερα ενοχλητικά".

Συμπερασματικά

Περιληπτικά, παρουσιάστηκε μια προσέγγιση για την ταυτοποίηση (εντοπισμό και εκτίμηση) εντοπισμένων, διαγωνίων (δηλαδή χωρίς εισδοχές/απαλοιφές), συστοιχίσεων αλληλουχιών, βασισμένη σε δραστικές προεκτάσεις της κλασσικής ιδέας των διαγωνίων διαγραμμάτων. Επειδή ελέγχεται διαρκώς, στενά και άμεσα η σημαντικότητα των όλο και μεγαλύτερων συνενούμενων συστοιχίσεων, δεν χρειάζονται εξωτερικές παρεμβάσεις, αρκεί να υπάρχει ένας κανόνας για τον ορισμό του σημείου εκκίνησης, δηλαδή των αρχικών σημειακών συστοιχίσεων, και μια μαθηματικά διατυπωμένη σημαντικότητα (όχι κατ'ανάγκη στατιστική) οποιασδήποτε αυθαίρετης συστοίχισης. Η προσέγγιση είναι προσαρμόσιμη σε διάφορα προβλήματα, ενώ εξωτερικά φίλτρα -αν και δεν απαιτούνται- μπορούν να αυξήσουν την αποτελεσματικότητα ή να βοηθήσουν την εξαγωγή διαφόρων εξειδικευμένων αποτελεσμάτων.

Η παρούσα προσέγγιση δεν αναζητά το τμήμα με τη μέγιστη αθροιστική ομοιότητα (maximal segment pair) όπως γίνεται στην τεχνική των Altschul και συνεργατών [Karlin και Altschul, 1990· Altschul et al, 1990]. Ούτε εκτελεί συστηματικό έλεγχο όλων των συνδυασμών παραθύρου/βαθμολογίας για να διαλέξει το καλύτερο, όπως στην περίπτωση του Argos [1987α]. Ούτε δοκιμάζει όλες τις πιθανές υπο-συστοιχίσεις σε κάθε διαγώνιο για να βρεί την καλύτερη όπως στην περίπτωση των Altschul και Erickson [1986]. Δεν θα ενώσει δύο υπο-συστοιχίσεις που βρίσκονται στην ίδια διαγώνιο, αλλά χωρίζονται από μια τρίτη που για κάποιο λόγο δεν μπορεί να ενωθεί με καμιά από τις δύο, έστω κι αν η τριάδα έχει συνολικά καλύτερη σημαντικότητα, γιατί οι άμεσα γειτονικές υπο-συστοιχίσεις πρέπει να ενωθούν πρώτες. Δεδομένου ενός κατάλληλου ορισμού για το σημείο εκκίνησης, αυτό το φαινόμενο μπορεί να βοηθήσει στον εντοπισμό των περιοχών μιας συστοίχισης που πραγματικά έχουν αξία, δηλαδή περιοχές υψηλής ομοιότητας ή απρόσμενης διαφοράς. Βλέπουμε για άλλη μια φορά ότι είναι η ίδια η φιλοσοφία της προσέγγισης που φαίνεται να αντιστοιχεί με πιο φυσικό τρόπο στις διαδικασίες που είναι κρυμμένες πίσω από τις παρατηρούμενες ομοιότητες και διαφορές.

Αναφορές κεφαλαίου

Altshul, S.F. (1991) "Amino Acid Substitution Matrices from an Information Theoretic Perspective" *J Mol. Biol.*, **219**, 555-565

Altshul, S.F. and Erickson, B.W (1986) "A Non-Linear Measure of Subalignment Similarity and its Significance Levels" *Bull. Math. Biol.*, **48**, 617-632

Altshul, S.F., Gish, W., Miller, W., Mayers, E.W. and Lipman, D.J. (1990) "Basic Local Alignment Search Tool" *J. Mol. Biol.*, **215**, 403-410

Argos, P. (1987α) "A Sensitive Procedure to Compare Amino Acid Sequences" *J Mol. Biol.*, **193**, 385-396

Argos, P. (1987β) "Analysis of Sequence-Similar Pentapeptides in Unrelated Tertiary Structures. Strategies for Protein Folding and a Guide for Site-Directed Mutagenesis" *J Mol. Biol.*, **197**, 331-348

Dayhoff, M. (1978) "Atlas of protein sequence and structure" Vol. **5**, suppl. 3, Washington D.C. : National Biomedical Research Foundation, pp. 345-358

Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) "Exhaustive Matching of the Entire Protein Sequence Database" *Science*, **256**, 1443-1445

Henikoff, S. and Henikoff, J G. (1992) "Amino Acid Substitution Matrices from Protein Blocks" *PNAS*, **89** 10915-10919

Henikoff, S. and Henikoff, J.G. (1993) "Performance Evaluation of Amino Acid Substitution Matrices" *Proteins: Struc. Fun. Gen.*, **17**, 49-61

Karlin, S. and Altshul, S.F. (1990) "Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes" *PNAS*, **87**, 2264-2268

Lipman, D.J. and Pearson, W R. (1985) "Rapid and Sensitive Protein Similarity Searches" *Science*, **227**, 1435-1441

Needleman, S.B. and Wunsch, C.D. (1970) "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins" *J . Mol. Biol.*, **48**, 443-453

Pearson, W.R. and Lipman, D.J. (1988) "Improved Tools for Biological Sequence Comparison" *PNAS*, **85** 2444-2448

Sander, C. and Schneider, R. (1991) "Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment" *Proteins: Struc. Fun. Gen.*, **9**, 56-68

Smith, T.F. and Waterman, M.S. (1981) "Identification of Common Molecular Subsequences" *J Mol. Biol.*, **147**, 195-197

Waterman, M.S. and Vingron, M. (1994) “Rapid and Accurate Estimates of Statistical Significance for Sequence Data Base Searches” *PNAS*, **91**, 4625-4628

Wilbur, W.J. and Lipman, D.J. (1983) “Rapid Similarity Searches of Nucleic Acid and Protein Data Banks” *PNAS*, **80**, 726-730