Προσαρμογή Πρωτεϊνης σε Πλέγμα (*Fitting Protein Chains to Lattices*)

Λέξεις-Κλειδιά: υπολογιστική βιολογία, υπολογιστική γεωμετρία, αλγόριθμος, πολυπλοκότητα.

1 Εισαγωγή

΄Ενα πολύ σημαντικό πρόβλημα στη βιολογία είναι ο καθορισμός (πρόβλεψη) της τρισδιάστατης δομής μιας πρωτεϊνης εάν δίνεται η ακολουθία των αμινοξέων που την αποτελούν. Με άλλα λόγια, πόσο γρήγορα μπορούμε να υπολογίσουμε (προβλέψουμε) το σχήμα που παίρνει η πρωτεϊνη στο χώρο αν ξέρουμε την ακολουθία αμινοξέων που την αποτελούν;

Το πρόβλημα αυτό που είναι γνωστό σαν 'το πρόβλημα του διπλώματος της πρωτεϊνης' (*Protein Folding problem*) φαίνεται να είναι εξαιρετικά δύσκολο ακόμη και αν εξεταστεί σε πολύ απλά μοντέλα: για παράδειγμα αν περιοριστούμε στους τρόπους που μπορεί να διπλώσει μια πρωτεϊνη πάνω σε ένα πλέγμα (*lattice*) και λάβουμε υπόψιν μόνο τις υδροφοβικές αλληλεπιδράσεις μεταξύ των αμινοξέων που γειτονεύουν στο πλέγμα. ΄Εχει αποδειχθεί [1, 3] ότι το πρόβλημα είναι NP-complete ακόμη και σε αυτό το απλοποιημένο μοντέλο.

Παρά το γεγονός ότι το πρόβλημα του διπλώματος της πρωτεϊνης σε πλέγματα είναι NP-complete, έχουν βρεθεί προσεγγιστικοί αλγόριθμοι [6, 13] οι οποίοι σε πολυωνυμικό χρόνο επιστρέφουν λύσεις οι οποίες απέχουν από τις βέλτιστες μόνο κατά ένα σταθερό προσεγγιστικό παράγοντα. Φυσικά, ακόμη και αν βρεθεί το βέλτιστο (πραγματικό) δίπλωμα μιας πρωτεϊνης πάνω σε ένα συγκεκριμένο πλέγμα, το δίπλωμα αυτό μπορεί να είναι πολύ διαφορετικό από το πραγματικό δίπλωμα της πρωτεϊνης στο χώρο. Συνεπώς η αναγνώριση 'καλών' μοντέλων πλέγματος που έχουν τη δυνατότητα να αναπαριστούν διπλώματα πρωτεϊνών που είναι πολύ κοντά στις πραγματικές τρισδιάστατες δομές τους είναι ένα σημαντικό πρόβλημα με το οποίο έχουν ασχοληθεί αρκετοί ερευνητές [2, 7, 5, 18, 17, 14, 10, 16, 9, 12].

Για να μετρήσουμε την ακρίβεια της αναπαράστασης διαφόρων μοντέλων πλέγματος χρησιμοποιούμε συνήθως την παρακάτω διαδικασία:

(1) Διάλεξε ένα σύνολο (*test set*) από πρωτεϊνες με γνωστή δομή στο χώρο και ένα σύνολο από διαφορετικά μοντέλα πλέγματος.

(2) Για κάθε ζεύγος πρωτεϊνης-πλέγματος βρες τη βέλτιστη αναπαράσταση της συγκεκριμένης πρωτεϊνης στο συγκεκριμένο πλέγμα, ελαχιστοποιώντας την συνολική απόσταση της αναπαράστασης στο πλέγμα από την πραγματική δομή στο χώρο. Αυτή η απόσταση συνήθως μετριέται με τη μέση τετραγωνική απόκλιση (*root mean squared deviation (RMS)*) των συντεταγμένων (*c-RMS*) ή των αποστάσεων (*d-RMS*).

(3) Για κάθε διαφορετικό πλέγμα υπολόγισε τη μέση τιμή των c-RMS (ή d-RMS) πάνω σε όλες τις πρωτεϊνες του test-set. Το πλέγμα με την μικρότερη μέση τιμή c-RMS (ή d-RMS) αναπαριστά καλύτερα τις πρωτεϊνες του test-set.

Το κρίσιμο μέρος στην παραπάνω διαδικασία είναι το βήμα (2), δηλαδή ο υπολογισμός της βέλτιστης αναπαράστασης μιας πρωτεϊνης σε ένα πλέγμα με δεδομένη την αναπαράσταση της πρωτεϊνης στο χώρο. Αυτό το πρόβλημα είναι γνωστό στη βιβλιογραφία σαν *Protein Chain Lattice Fitting (PCLF) problem*. Επίσης το βρίσκουμε σαν '*the discretization of a protein backbone*', '*modeling protein structures on a lattice*', '*lattice approximation of 3D structure of a chain molecule*', '*discrete state model fitting to X-ray structures*', κλπ., [2, 16, 9, 14, 12, 7, 17, 18, 11, 5].

Πρόσφατα αποδείχθηκε ότι το PCLF πρόβλημα είναι NP-complete σε τρισδιάστατα πλέγματα με πλευρά 3.8A χρησιμοποιώντας την απόκλιση c-RMS [11]. Ένα πλήθος από ευριστικούς (*heuristics*) και εκθετικούς αλγόριθμους έχει προταθεί για αυτό το πρόβλημα: Ένας από τους πρώτους αλγόριθμους που προτάθηκε στο [2], απαριθμεί όλες τις δυνατές αναπαραστάσεις και επιλέγει την καλύτερη. Αλγόριθμοι που βασίζονται σε δυναμικό προγραμματισμό έχουν παρουσιαστεί σε διάφορες εργασίες [18, 17, 16]. Ένας άπληστος (*greedy*) αλγόριθμος ο οποίος κρατά περίπου 500 'πολύ καλά' διπλώματα της πρωτεϊνης στο πλέγμα χρησιμοποιήθηκε στο [14] και ένας άλλος άπληστος αλγόριθμος παρουσιάστηκε στο [12]. Επίσης παρουσιάστηκαν αλγόριθμοι ακέραιου προγραμματισμού που δίνουν βέλτιστες λύσεις αλλά σε εκθετικό χρόνο στο [8]. Τέλος στο [15] οι συγγραφείς παρουσιάζουν έναν randomized αλγόριθμο (τύπου Monte-Carlo) για να προσεγγίσουν το βέλτιστο ταίριασμα μιας πρωτεϊνης σε τρισδιάστατο πλέγμα.

Όλοι οι παραπάνω αλγόριθμοι είτε είναι εκθετικής πολυπλοκότητας (που σημαίνει ότι μπορούν να μας δώσουν γρήγορα απαντήσεις μόνο για μικρές σε μήκος αλυσίδες πρωτεϊνών), είτε είναι προσεγγιστικοί χωρίς καμμιά όμως εγγύηση για το πόσο καλή είναι η λύση που παράγουν (δηλαδή χωρίς καμμιά θεωρητική απόδειξη για το πόσο απέχει η λύση τους από τη βέλτιστη λύση). Σημειώνουμε εδώ ότι το πρόβλημα λύνεται βέλτιστα σε πολυωνυμικό χρόνο αν περιοριστούμε σε μονοδιάστατο πλέγμα [4].

2 Τυπικός ορισμός του προβλήματος

Τυπικά το πρόβλημα PCLF μπορεί να οριστεί ως εξής:

Δεδομένα:

- Η αναπαράσταση μιας αλυσίδας αμινοξέων σαν ένα μονοπάτι σημείων στον τρισδιάστατο χώρο. Για κάθε σημείο δίνονται οι συντεταγμένες του στο χώρο. Το μονοπάτι φυσικά δεν τέμνει τον εαυτό του και καθορίζει τη διάταξη των αμινοξέων.
- Ένα πλέγμα L.

Ζητούμενο:

Μια απεικόνιση η οποία αντιστοιχίζει κάθε ένα σημείο του μονοπατιού (που αναπαριστά κάποιο στοιχείο της πρωτεϊνης) σε ένα σημείο του πλέγματος L έτσι ώστε η 'απόσταση' (π.χ., c-RMS, d-RMS, κλπ.) μεταξύ της αναπαράστασης της πρωτείνης στο χώρο και της εικόνας της στο πλέγμα να είναι ελάχιστη και να ισχύουν οι παρακάτω ιδιότητες:

- Δύο διαφορετικά σημεία του μονοπατιού αντιστοιχίζονται σε δύο διαφορετικά σημεία του πλέγματος L.
- Δύο διαδοχικά σημεία του μονοπατιού αντιστοιχίζονται σε δύο γειτονικά σημεία του πλέγματος L.

3 Ανοιχτά προβλήματα - Προτεινόμενη έρευνα

Τα παρακάτω προβλήματα έχουν παραμείνει ανοιχτά στο [11].

Από θεωρητική άποψη είναι πολύ ενδιαφέρουσα η έρευνα της πολυπλοκότητας του προβλήματος σε δισδιάστατα πλέγματα. Η απόδειξη της NP-πληρότητας που παρουσιάστηκε στο [11] (για την οποία χρησιμοποιήθηκε μία αναγωγή από μία ειδική περίπτωση του προβλήματος ικανοποιησιμότητας (SAT)) δεν μπορεί να εφαρμοστεί στην περίπτωση του δισδιάστατου πλέγματος. Έτσι παραμένει ανοιχτό το πρόβλημα της ύπαρξης αλγόριθμου πολυωνυμικού χρόνου για αυτήν την περίπτωση.

Από πρακτική άποψη τα παρακάτω προβλήματα παραμένουν ανοιχτά:

- Ποια είναι η πολυπλοκότητα του προβλήματος σε διαφορετικούς τύπους τρισδιάστατων πλεγμάτων (π.χ., ανισόπλευρα, μή-ορθογώνια, κλπ.).

- Ποια είναι η πολυπλοκότητα του προβλήματος αν χρησιμοποιηθεί η d-RMS απόκλιση σαν κριτήριο μέτρησης της απόστασης μεταξύ των δύο τρισδιάστατων δομών.

Είναι επίσης ενδιαφέρον να μελετηθεί αν κάποιος από τους υπάρχοντες προσεγγιστικούς αλγόριθμους έχει εγγύηση καλής προσέγγισης (θεωρητική απόδειξη) ή έστω φαίνεται να συγκλίνει σε μια καλή προσέγγιση (πειραματική μελέτη), καθώς και να προταθεί καινούριος αλγόριθμος με κατά προτίμηση σταθερό παράγοντα προσέγγισης.

Τέλος είναι ενδιαφέρουσα η πειραματική σύγκριση της απόδοσης των γνωστών αλλά και καινούριων αλγόριθμων όταν αυτοί εφαρμοστούν σε συγκεκριμένα σύνολα πρωτεινών.

Βιβλιογραφία

[1] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J. Comp. Biol., 5:27-40, 1998.

[2] D. G. Covell and R. L. Jernigan. Conformations of folded proteins in restricted spaces. Biochemistry, 29:3287-3294, 1990.

[3] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. In Proc. of STOC '98, pp. 597-603, 1998.

[4] I. Emiris, E. Kranakis, E. Markou, E. Tsigaridas. Finding the best on-lattice fit, manuscript.

[5] A. Godzik, A. Kolinski, and J. Skolnick. Lattice representations of globular proteins: How good are they? J. Comp. Chem., 14:1194-1202, 1993.

[6] W.E. Hart, S. Istrail, Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal, J. Comput. Biol. 3, 53-96, 1996.

[7] D. A. Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. USA, 89:2536–2540, 1992.

[8] X. Huang. Fitting protein chains to lattice using integer programming approach. M.Sc. thesis. School of Computing Science, Simon Fraser University, 2007.

[9] P. Koehl and M. Delarue. Building protein lattice models using self-consistent mean field theory. J. Chem. Physics, 108(22):9540–9549, 1998.

[10] R. Kolodny, P. Koehl, L. Guibas and M. Levitt, Small Libraries of Protein Fragments Model Native Protein Structures Accurately, J. Mol. Biol. 323, 297–307, 2002.

[11] J. Manuch and D. Gaur. Fitting protein chains to cubic lattice is NP-complete. J. Bioinform. Comput. Biol., 6:93–106, 2008.

[12] C. Mead, J. Manuch, X. Huang, B. Bhattacharyya, L. Stacho, and A. Gupta. Investigating lattice structure for inverse protein folding (poster abstract). FEBS Journal, 272 (s1):4739 1 380, 2005.

[13] A. Newman, A new algorithm for protein folding in the HP model, in Proceedings of the 13th ACM- SIAM, Symposium on Discrete Algorithms, 876–884, 2002.

[14] B. H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol., 249:493–507, 1995.

[15] Y. Ponty, R. Istrate, E. Porcelli, P. Clote. LocalMove: Computing on-lattice fits for biopolymers, Nucleic Acids Research, Vol. 36, Web Server issue, 2008.

[16] A. A. Rabow and H. A. Sheraga. Improved genetic algorithm for the protein folding problem by use of a Cartesian combination operator. Protein Science, 5:1800–1815, 1996.

[17] B. A. Reva, D. S. Rykunov, A. J. Olson, and A. V. Finkelstein. Constructing lattice models of protein chains with side groups. Journal of Comp. Biology, 2(4):527–535, 1995.

[18] D. S. Rykunov, B. A. Reva, and A. V. Finkelstein. Accurate general method for lattice approximation of three-dimensional structure of a chain molecule. Proteins: Structure, Function and Genetics, 22:100–109, 1995.