Informatica Biomedica

lezione15

Alberto*Paoluzzi Mauro*Ceccanti
http : //www.dia.uniroma3.it/ paoluzzi/web/did/biomed/

Informatica e Automazione, "Roma Tre" — Medicina Clinica, "La Sapienza"

May 17, 2010

Informatica Biomedica: Lezione 15

Macromolecular structures

Superposition of structures, and structural alignments

Algoritmo BioEuler
- Metodo di calcolo dell'indice
- Metodo di Confronto tra due strutture

Punti critici dell'approccio
- Dimensione relativa delle due biomolecole
- Bipartizione in parti uguali

Segmentazione dei domini strutturali

Suddivisione basata sulla struttura secondaria?

The Worldwide Protein Data Bank (wwPDB)

The World Wide PDB (wwPDB) is a collaboration between three primary archival projects to integrate the archiving and distribution of biological macromolecular structures

- The Research Collaboratory for Structural Bioinfornatics (RCSB) (USA)
- The EBI Protein Structure Database in Europe or Macromolecular Structure Database (MSD) (at the European Bioinformatics Institute (EBI), Hinxton. UK)
- The Protein Data Bank/Japan (Osaka, Japan)

Other structure databanks

Other databanks reorganize and provide access to the data, including:

- Structural Classification of Proteins (SCOP) is a carefully curated database of all protein domains, classified according to structure, function and evolution.
- The Molecular Modeling DataBase (MMDB) is the project within the US National Center for Biotechnology Information (NCBI) ENTREZ system, treating experimentally determined macrolll01ecular structures.

Structural comparison by RMSD

The average distance between corresponding points is a measure of the structural similarity.

- In practice it is conventional to compute the root-mean-square deviation (RMSD) of the corresponding atoms:
- root-mean-square deviation

$$RMSD = \sqrt{\frac{\sum_i d_i^2}{n}}$$

Example of sequence aligment

Figure: sequence aligment

```
8gch CGVPAIQPVLIVNG----------------------------EEAVP--GS----WPWQVSLQ-DKTG
1agj --------------EVSAEEIKKHEEKWNKYYGVNAFNLPKELFSKVDEKDR-QKYPYNTIGNVFVK-G-

8gch FH--FCGGSLINE-NWVVTAAHC-GV-T---T-SDVVVAGEFDQG---SSSEKI--QKLKIAKVFK-NS-
1agj --QTSATGVLIG-KNTVLTNRHIAK-FANGDPSKVSFRPSI-NTDDNGNT-E-TPYGEYEVKEILQEP-F

8gch KYNSLTINNDITLLKLST-----AAS--FSQTVSAVCLPSASD--DFAAGTTCVTTGWG-LTRYNTPD-R
1agj GAG-----VDLALIRLKPDQNGVSL-GDK---ISPAKIGT---SNDLKDGDKLELIGYPFDH----KVNQ

9gch LQQASLPLL-SNTNCKKYWGTKIKDAM--ICAGASGV-SSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTC
1agj MHRSEIELTTLS--------------RGLRYY----GFTVPGNSGSGIFNSN---GELVGIHSSK----

8gch STST---------PGVYARVTA-LVNWVQQTLAAN-
1agj ----VSHLDREHQINYGVGIGNYVKRIINEKN---E
```

Example of Structural superposition

Structural superposition of y-chymotrypsin 8GCH (black) and S. aureus epidermolytic toxin A 1AGJ (blue)

Goal of the BioEuler algorithm

Allineamento strutturale basato su un indice tensoriale gerarchico.

Albero binario di tensori del secondo ordine (trasformazioni affini).

Algoritmo

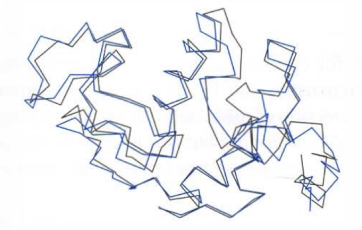
- allineamento della struttura sul sistema principale
- calcolo della matrice diagonale corrispondente
- bisezione della struttura
- calcolo del tensore di Eulero esteso per le due parti
- bisezione delle parti
- ripetizione ricorsiva del calcolo per n livelli ...

Confronto di due strutture

- k = 1
- calcolo dell'indice di Eulero di livello k di entrambe
- se la distanza euclidea dei due indici è minore di una soglia
  - calcolo dell'indice di livello k+1
  - ripeti dal passo 2
- altrimenti termina

Example of structural alignment

Exercise
Compute the Euler indices of two related sequences.
Compare with RMSD



- Aligned sequences, and superposed structures, of two related proteins
  1. egg white lysozyme (black)
  2. baboon α-lactalbumin (blue)

Dimensione delle due biomolecole

Non sembra ragionevole confrontare molecole di dimensioni (numero di atomi) significativamente differenti.

Il metodo sembra però ben fondato per molecole di dimensioni simili, in particolare per determinare quali molecole debbano essere considerate appartenenti alla stessa famiglia; in secondo luogo per creare un grafo pesato di similarita' tra le molecole di una stessa famiglia.

In particolare, questo approccio di confronto gerarchico basato su indici globali sembra prestarsi benissimo allo screening completo di un database, per organizzarlo in famiglie e sottofamiglie strutturali di forma (e dimensione) simile.

Pertanto sembra particolarmente utile per una ri-classificazione gerarchica dei domini funzionali (ad esempio estratti dal database SCOP).

Bipartizione in parti uguali

Anche se la biratizione in parti uguali potrebbe sembrare arbitraria, in quanto non suddivide in corrispondenza dei bordi dei domini funzionali, nondimeno consente un confronto efficace sulla forma globale della molecole sulla distribuzione spaziale, anche fine, delle masse atomiche e sulla disposizione locale dei siti funzionali.

A questo scopo sembra opportuna una normalizzazione dell'indice, che filtri l'effetto di piccole variazioni di numerosità (o di peso) degli atomi (o dei residui) della molecola, visto che ogni atomo apporta il contributo additivo del suo tensore di Eulero all'indice in costruzione. Questo può essere facilmente normalizzato dividendo per il termine di massa, in modo tale da ottenere sempre matrici euleriane con termine [4,4] eguale ad 1.

Domain definitions

"The definition of protein domains varies widely across the discipline of biology. Domains are defined simultaneously as:

(1) regions that display a significant level of sequence homology;
(2) a minimal part of the gene that is capable of performing a function;
(3) a region of the protein with an experimentally assigned function;
(4) parts of structures that have significant structural similarity; and
(5) compact spatially distinct units of protein structure."

[Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN.
Toward consistent assignment of structural domains in proteins.
J Mol Biol. 2004 Jun 4;339(3):647-678.]

Segmentazione di una struttura: estrazione dei domini

Per isolare i domini 3D costituenti una proteina, bisogna innanzitutto avere una chiara definizione del concetto di dominio, come illustrato nelle slide successive.

STRUDL (STRUctural Domain Limits)

[Wernisch L, Hunting M, Wodak SJ. Identifi cation of structural domains in proteins by a graph heuristic. Proteins. 1999 May 15;35(3):338-352.]

Algorithm designed to identify domains with any number of non-contiguous chain segments.

Uses the Kernighan-Lin graph heuristic to partition the protein into residue sets which display minimum interactions between them.

Interactions are deduced from contact areas between atoms in the weighted Voronoi diagram.

The radius of the "accessible sphere" around each atom is the van der Waals radius of the atom increased by $1.4 angstrom$.

DOMAK (DOmain MAKer)

[Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain defi nitions. Protein Sci. 1995 May;4(5):872-884.]

Based on the principle that the residues comprising a domain make more contacts between themselves (internal contacts) than they do with the rest of the protein (external contacts).

Two residues make contact if a heavy atom of one is within $5 angstrom$ of a heavy atom of the other.

$$\frac{int_A}{ext_{AB}} \times \frac{int_B}{ext_{AB}}$$

Can deal with domains consisting of two segments.

Conflicting domain assignments

"The major factors responsible for conflicting domain assignments between methods, both experts and automatic, are:

(1) the defi nition of very small domains;
(2) splitting secondary structures between domains;
(3) the size and number of discontinuous domains;
(4) closely packed or convoluted domain-domain interfaces;
(5) structures with large and complex architectures; and
(6) the level of signifi cance placed upon structural, functional and evolutionary concepts in considering structural domain defi nitions."

[Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN.
Toward consistent assignment of structural domains in proteins.
J Mol Biol. 2004 Jun 4;339(3):647-678.]

Clustering sulle strutture secondarie

Un approccio alternativo a quello top-down illustrato in precedenza potrebbe essere uno di tipo aggregativo bottom-up, che aggregasse gerarchicamente i domini a partire da unità strutturali quali i componenti della struttura secondaria, e che sono chiaramente delineati dal punto di vista biochimico.

In altri termini i domini stessi potrebbero essere definiti come una sorta di superstruttura secondaria, attraverso tecniche di clustering sulle strutture secondarie (α-eliche e β-sheets).