Lecture 9

Predicting domain features from sequence

Fundamentals of Bioinformatics

Protein Domain delineation

Content:
- Background: what is a domain?
- Predicting domains from structure
- Predicting domains from sequence alone
  - Linker prediction (DomCut, Elsik)
  - Protein domain delineation based on consistency of multiple ab initio model tertiary structures (SnapDRAGON, Rosetta)
  - Protein domain delineation based on combining homology searching with domain prediction (Domaination)
  - Domain delineation based on sequence hydrophobicity patterns (SCOOBY-Domain)

A protein domain is a:

- **Structural**: Compact, semi-independent unit (Richardson, 1981)
- **Stable unit of a protein structure that can fold autonomously** (Wetlaufer, 1973)
- **Functional**: Fundamental unit of protein function
- **Evolutionary**: Recurring functional and evolutionary module (Bork, 1992)

“Nature is a ‘tinkerer’ and not an inventor” (Jacob, 1977).

diphtheria toxin (1ddt)

A beautiful example of a multi-domain structure is diphtheria toxin, made up of three domains, each of which is involved in a different stage of infection (receptor binding, membrane penetration, and catalysis of ADP-ribosylation of elongation factor 2). A structural neighbor is depicted next to each domain of diphtheria toxin (middle).

Domain characteristics

- Domains are genetically mobile units, and multidomain families are found in all three kingdoms (Archaea, Bacteria and Eukarya)
- The majority of genomic proteins, 75% in unicellular organisms and more than 80% in metazoa, are multidomain proteins created as a result of gene duplication events (Apic et al., 2001).
- Domains in multidomain structures are likely to have once existed as independent proteins, and many domains in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes (Davidson et al., 1993).

The DEATH Domain

- Present in a variety of Eukaryotic proteins involved with cell death.
  - Six helices enclose a tightly packed hydrophobic core.
  - Some DEATH domains form homotypic and heterotypic dimers.

A beautiful example of a multi-domain structure is diphtheria toxin, made up of three domains, each of which is involved in a different stage of infection (receptor binding, membrane penetration, and catalysis of ADP-ribosylation of elongation factor 2). A structural neighbor is depicted next to each domain of diphtheria toxin (middle).

http://www.mshri.on.ca/pawson
Delineating domains is essential for:

• Obtaining high resolution structures by NMR (due to size limitations of proteins)
• Sequence analysis
  • Multiple sequence alignment methods
  • Prediction algorithms (secondary/tertiary structure, solvent accessibility, …)
• Fold recognition and threading
• Structural/functional genomics
• Cross genome comparative analysis
• Elucidating the evolution, structure and function of a protein family (e.g. ‘Rosetta Stone’ method)

Prediction of protein-protein interactions

Rosetta stone

- Gene fusion is an effective method for prediction of protein-protein interactions
- If proteins A and B are homologous to two domains of a protein C, A and B are predicted to have interaction

Though gene-fusion has low prediction coverage, its false-positive rate is low

Domain fusion example

- Vertebrates have a multi-enzyme protein (GARs-AIRs-GART) comprising the enzymes GAR synthetase (GARs), AIR synthetase (AIRs), and GAR transformylase (GART).
- In insects, the polypeptide appears as GARs-(AIRs)2-GART.
- In yeast, GARs-AIRs is encoded separately from GART.
- In bacteria each domain is encoded separately (Henikoff et al., 1997).

GAR: glycinamide ribonucleotide
AIR: aminomimidazole ribonucleotide

Structural domain organisation can be nasty

Pyruvate kinase
Phosphotransferase

1 continuous + 2 discontinuous domains

Domain connectivity

A continuous domain is often an evolutionary module

Domain size

- The size of individual structural domains varies widely
  • from 36 residues in β-selectin to 692 residues in lipoygenase-1 (Jones et al., 1998)
  • the majority (90%) having less than 200 residues (Siddiqui and Barton, 1995)
  • with an average of about 100 residues (Islam et al., 1995).
- Small domains (less than 40 residues) are often stabilised by metal ions or disulphide bonds.
- Large domains (greater than 300 residues) are likely to consist of multiple hydrophobic cores (Garel, 1992).
**Delineating Structural Domains**

- A structural domain may be detected as a compact, globular substructure with more interactions within itself than with the rest of the structure (Janin and Wodak, 1983).
- Therefore, a structural domain can be determined by two shape characteristics: compactness and its extent of isolation (Tsai and Nussinov, 1997).
- Measures of local compactness in proteins have been used in many of the early methods of domain assignment (Rossmann et al., 1974; Crippen, 1978; Rose, 1979; Go, 1978) and in several of the more recent methods (Holm and Sander, 1994; Islam et al., 1995; Siddiqui and Barton, 1995; Zehfus, 1997; Taylor, 1999).

**Detecting Structural Domains**

Protein core is densely packed

Contact plot

A protein domain has many residue contacts within a domain (intra-domain contacts) and relatively few between domains (inter-domain contacts)

**Detecting Structural Domains**

- Approaches encounter problems when faced with highly associated domains (and sometimes also with discontinuous ones) and many definitions will require manual interpretation.
- Consequently there are discrepancies between assignments made by domain databases (Hadley and Jones, 1999).

**Detecting Structural Domains**

More recent methods are better:

- Taylor (1999)

**Detecting Structural Domains**

Taylor method (1999)

**DOMAIN-3D**

- Easy and clever method
- Uses a notion of spin glass theory (disordered magnetic systems) to delineate domains in a protein 3D structure
- Steps:
  1. Take sequence with residue numbers (1..N)
  2. Look at neighborhood of each residue (first shell)
  3. If ("average nghhood residue number" > res no)
     resno = resno+1
     else resno = resno-1
  4. If (convergence) then take regions with identical "residue number" as domains and terminate

When a structural template is available:

**Modelling by homology**
- Find template structure
- Align query sequence onto template sequence
- Model the backbone
- Fill the gaps by loop modelling
- Build in the side-chains (e.g. copy conserved side-chains and model mutated ones)

But: alignment of multi-domain proteins can be difficult

Predicting domain boundaries from linker regions
- Needed: discernible signal that sets linker regions apart from other sequence regions
- Problems:
  - Linker regions are short, difficult to get statistical signal
  - Linker regions versus intra-domain loops, not much statistical difference
  - No distinction continuous/discontinuous domain possible

Detecting Domains using Sequence only
- Even more difficult than prediction (i.e. delineation) from structure!
- In this lecture we will limit ourselves to prediction of domain boundaries
  - so just in between which pair of residues is a domain split?
- Any ideas?

Repeat for all amino acids until convergence

if \( 41 < \frac{5+56+78+89}{5} \)
then Res 41–42 (up 1)
else Res 41–40 (down 1)

Domain-3D: output

Convergence of residue numbering for 2-domain protein with discontinuous and continuous domain

Convergence of residue numbering for continuous 2-domain protein

Initial situation with ‘normal’ residue numbers 1..N

Taylor method (1999)

1, 2, 3, ..., 198, 199, 200

49, 49, ..., 49, 151, ..., 151, 151

Detecting Domains using Sequence only

- Even more difficult than prediction (i.e. delineation) from structure!
- In this lecture we will limit ourselves to prediction of domain boundaries
  - so just in between which pair of residues is a domain split?
- Any ideas?
**Predicting domain boundaries from linker regions - approaches:**

- Building linker index (using amino-acid propensities for being within linker or non-linker - using known structures):
  - LinkerDB (George & Heringa, 2002)
  - Domcut (Suyama & Ohara, 2003) - Sens./Spec. ~ 50%

\[ S_i = -\log \left( \frac{p_{\text{linker}}}{p_{\text{domain}}} \right) \]

where \( i \) denotes the amino acid type and \( f \) the frequencies in either linker or domain

\[ L-L, L-D, D-D, D-L \text{ transitions} \]

\[ \text{L = linker, D = domain} \]

**SnapDRAGON**

- **Scientific Name**
  Antirrhinum majus

- **Common Name**
  Snapdragon

Sequence searches using PSI and alignment by PRALINE (Heringa, 1999) 
Sum proposed boundary positions and predicted secondary structure 
Generate 100 DRAGON 3D models 
Assign domain boundaries

PREDATOR secondary structure prediction program
Convert to Z
followed by

SNAPDRAGON 
Domain boundary prediction protocol using sequence information alone (Richard George)

1. Input: Multiple sequence alignment (MSA) and predicted secondary structure 
2. Generate 100 DRAGON 3D models of the protein structure associated with the MSA 
3. Assign domain boundaries to each of the 3D models (Taylor, 1999) 
4. Sum proposed boundary positions within 100 models along the length of the sequence, and smooth boundaries using a weighted window
   - Convert to Z-scores

These four steps will be explained on the next slides


SNAPDRAGON 
Domain boundary prediction protocol using sequence information alone (Richard George)

STEP 1: 
Input: Multiple sequence alignment (MSA) 
1. Sequence searches using PSI-BLAST (Altschul et al., 1997) 
2. followed by sequence redundancy filtering using OBSTRUCT (Heringa et al., 1992) 
3. and alignment by PRALINE (Heringa, 1999)
   - and predicted secondary structure
4. PREDATOR secondary structure prediction program


Information content of a multiple alignment
Align homologous sequences (ideally orthologues)
SNAPDRAGON
Domain boundary prediction protocol using sequence information alone (Richard George)

STEP 2:
Generate 100 DRAGON (Aszodi & Taylor, 1994) models for the protein structure associated with the MSA
- DRAGON folds proteins based on the requirement that (conserved) hydrophobic residues cluster together
- (Predicted) secondary structures are used to further estimate distances between residues (e.g. between the first and last residue in a β-strand).
- Based on these constraints, it compiles a target matrix with “desired” distances
- It then constructs 100 random high dimensional Cα (and pseudo Cβ) distance matrices
- For each distance matrix, distance geometry is used to find the 3D conformation corresponding to the prescribed target matrix of desired distances between residues (by gradual inertia projection and based on input MSA and predicted secondary structure)

DRAGON = Distance Regularisation Algorithm for Geometry OptimisatioN

STEP 3.
Assign domain boundaries to each of the 3D models (Taylor, 1999)
- See earlier slides

Taylor method (1999)
DOMAIN-3D
STEP 4

Sum proposed boundary positions within 100 models along the length of the sequence, and smooth boundaries using a weighted window (assign central position)

Window score = \( \sum_{i=1}^{l} S_i \times W_i \)

Where \( W_i = (p \cdot |p-j|)/p^2 \) and \( p = \frac{1}{2}(i+1) \).

It follows that \( \sum W_i = 1 \)

---

SnapDRAGON prediction assessment

- Test set of 414 multiple alignments; 183 single and 231 multiple domain proteins.
- Boundary predictions are compared to the region of the protein connecting two domains (maximally \( \pm 10 \) residues from true boundary)
  - Many methods use \( \pm 20 \) from true boundary… (what do you think about that?)

---

Average prediction results per protein

<table>
<thead>
<tr>
<th>Method</th>
<th>Coverage (\pm SD)</th>
<th>Success (\pm SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNAPDRAGON</td>
<td>63.9 (\pm 43.0)</td>
<td>46.8 (\pm 36.6)</td>
</tr>
<tr>
<td>Baseline 1</td>
<td>44.4 (\pm 25.5)</td>
<td>23.1 (\pm 27.3)</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>35.7 (\pm 41.3)</td>
<td>22.7 (\pm 27.3)</td>
</tr>
</tbody>
</table>

Coverage is the % linkers predicted (TP/TP+FN)
Success (PPV) is the % of correct predictions made (TP/TP+FP)

---

SnapDRAGON prediction assessment

- Baseline method I:
  - Divide sequence in equal parts based on number of domains predicted by SNAPDRAGON
- Baseline method II:
  - Similar to Wheelan et al., based on domain length partition density function (PDF)
  - PDF derived from 2750 non-redundant structures (deposited at NCBI)
  - Given sequence, calculate probability of one-domain, two-domain, … protein
  - Highest probability number of domains taken and sequence split equally as in baseline method I
SnapDRAGON
- Uses consistency in the absence of standard of truth
  - Models not need to be correct as long as most have proper domain boundaries (linker regions)
  - Goes from primary-secondary to tertiary structure to ‘just’ chop protein sequences
  - Is very slow (can be hours for proteins>400 aa) – need cluster or GRID implementation
  - SnapDRAGON webserver is underway
  - The SnapDRAGON protocol is used by the Baker group (UW, Seattle) in their RosettaDOM server (next slides)

RosettaDOM
- We developed a de novo domain prediction method that is similar in concept to SnapDRAGON but uses the Rosetta de novo structure prediction method to produce models.
  - RosettaDOM generates 400 three-dimensional models using Rosetta, and then selects the top 200 scoring models that pass filters that eliminate structures with too many local contacts or unlikely strand topologies.
  - Domain boundaries are then assigned for each of the 200 models using Taylor’s structure-based domain-identification algorithm described above.
  - Final domain-boundary predictions are made based on consistencies found in the domain assignments of these models by taking the sum of boundary assignments at each position along the protein chain, smoothing the values using a center weighted sliding window, and then converting the smoothed boundary distributions to Z scores as described by George et al.15 Positions with Z scores of 2.5 or greater are treated as potential domain boundaries.
  - Because logic is not applied to assign discontinuous domains and continuous domains are unlikely to be less than 50 residues in length, final domain boundaries are assigned for positions with the highest Z scores that are at least 50 residues apart and are not within 50 residues of the N and C termini.

Integrated protein sequence database searching and on-the-fly domain recognition

Delineating domains using sequence database searches

DOMAINATION
Richard A. George

Protein domain identification and improved sequence searching using PSI-BLAST


Domaination
- Current iterative homology search methods (e.g. PSI-BLAST) do not take into account (that):
  - Domains may have different ‘rates of evolution’.
  - Common conserved domains, such as the tyrosine kinase domain, can obscure weak but relevant matches to other domain types
  - Premature convergence (false negatives)
  - Matrix migration / Profile wander (false positives).

DOMAINATION is based upon PSI-BLAST
query

Strategy: Combine C- and N-termini of local alignments to delineate domain boundaries
PSI (Position Specific Iterated) BLAST

- basic idea
- use results from BLAST query to construct a profile matrix
- search database with profile instead of query sequence
- iterate

PSI BLAST: Constructing the Profile Matrix

Figure from: Altschul et al. Nucleic Acids Research 25, 1997

PSI-BLAST iteration

- Query sequences are first scanned for the presence of so-called low-complexity regions (Wooton and Federhen, 1996 – next slide) which are masked
- The program then initially operates on a single query sequence by performing a gapped BLAST search
- Then, the program takes significant local alignments (hits) found, constructs a ‘multiple alignment’ (master-slave alignment) and abstracts a position-specific scoring matrix (PSSM) from this alignment.
- The database is rescanned in a subsequent round, using the PSSM, to find more homologous sequences. Iteration continues until user decides to stop or search has converged

PSI-BLAST steps in words

For reference only

A Profile Matrix (Position Specific Scoring Matrix - PSSM)

This is the same as a profile without position-specific gap penalties

For reference only
Low-complexity sequences

- For example: AAAAA… or AYLAYLAYL… or AYLLYAALY…
- Low-complexity (sub)sequences have a biased composition and contain less information than high-complexity sequences
- Because of the low information content, they often lead to spurious hits without a biological basis (for example, you can’t tell whether a poly-A sequence is more similar to a globin, an immunoglobulin or a kinase sequence).

PSI-BLAST entry page

1. This portion of each description links to the sequence record for a particular hit.
2. Score or bit score is a value calculated from the number of gaps and substitutions associated with each aligned sequence. The higher the score, the more significant the alignment. Each score links to the corresponding pairwise alignment between query sequence and hit sequence (also referred to as subject or target sequence).
3. E Value (Expect Value) describes the likelihood that a sequence with a similar score will occur in the database by chance. The smaller the E Value, the more significant the alignment. For example, the first alignment has a very low E value of e-117 meaning that a sequence with a similar score is very unlikely to occur simply by chance.
4. These links provide the user with direct access from BLAST results to related entries in other databases. ‘L’ links to LocusLink records and ‘S’ links to structure records in NCBI’s Molecular Modeling Database.

PSI-BLAST

A residues denote low-complexity sequence fragments that are ignored.
Identifying domain boundaries

1. Sum N- and C-termini of gapped local alignments
2. True N- and C-termini are counted twice (within 10 residues)
3. Boundaries are smoothed using two windows (15 residues long)
4. Combine scores using biased protocol:
   \[
   \text{if} \quad N \cdot C = 0
   \text{then} \quad S = N + C
   \text{else} \quad S = N \cdot C + (N + C)/2
   \]

Identifying domain deletions

- Deletions in the query (or insertion in the DB sequences) are identified by
  - two adjacent segments in the query align to the same DB sequences (>70% overlap), which have a region of >35 residues not aligned to the query.

Identifying domain permutations

- A domain shuffling event is declared
  - when two local alignments (>35 residues) within a single DB sequence match two separate segments in the query (>70% overlap), but have a different sequential order.

Creating domain profiles

- A representative set of the database sequence fragments that overlap a putative domain are selected for alignment using OBSTRUCT (Heringa et al. 1992).
  - >20% and <60% sequence identity (including the query seq).
- A multiple sequence alignment of these fragments is generated using PRALINE (Heringa 1999, 2002; Simossis et al., 2005).
- Each domain multiple alignment is used as a profile in further database searches using PSI-BLAST (Altschul et al 1997).
- The whole process is iterated until no new domains are identified.

Assessing Domaination boundary prediction accuracy

- Evaluated using a set of 452 multidomain proteins
  - 56% of proteins were correctly predicted to have more than one domain
  - 42% of predictions are within ±20 residues of a true boundary
  - 49.9% (±4.6%) correct boundary predictions per protein
Assessing Domaination boundary prediction accuracy (cont.)

• 23.3% of all linkers found in 452 multidomain proteins (low coverage). Not a surprise since:
  – Structural domain boundaries will not always coincide with sequence (motif) domain boundaries
  – Proteins must have some domain shuffling
• For discontinuous proteins 34.2% of linkers were identified (surprisingly high fraction)
• 30% of discontinuous domains were successfully joined (good for sequence only method)

Benchmarking sequence searching improvement versus PSI-BLAST

• A set 452 non-homologous multidomain protein structures
• Delineated each sequence using true structural domains
• Do PSI-BLAST database searches using individual domain sequences
• Tested to what extent PSI-BLAST and DOMAINATION, when run on the full-length protein sequences, can capture the sequences found by the reference PSI-BLAST searches using the individual domains.

SSEARCH significance test

• Verify the statistical significance of each database sequence found by relating it to the original query sequence (instead of to the PSSM created by PSI-BLAST at each iteration).
• SSEARCH (Pearson & Lipman 1988) was used. It calculates an E-value for each generated local alignment using Dynamic Programming (DP)
  • Aligns query and database sequences optimally
  • This filter will lose distant homologies (bad E-values).
  • It tests whether DOMAINATION and PSI-BLAST are able to capture DB sequences similar to the query
  • Use the 452 proteins with known structure.

Significant sequences found in database searches

At an E-value cut-off of 0.1 the performance of DOMAINATION searches with the full-length proteins is 15% better than PSI-BLAST over the full range of E-values

Scooby-domain: prediction of globular domains in protein sequence

Richard A. George1,2, Kuang Lin1 and Jaap Heringa4

1 Informatika Ltd, 60 Cheapside, London EC2V 6IT, UK
2 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
3 Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill NW7 1AA, UK
4 Centre for Integrative Bioinformatics (BBVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
4 Corresponding author

Generating a domain probability matrix for a query sequence

• Scooby-domain uses a multilevel smoothing window to predict the location of domains in a query sequence.

• Based on the window length and its average hydrophobicity, the probability that it can fold into a domain is found directly from the distribution of domain size and hydrophobicity, calculated using sequence-level domain representatives from the CATH domain database (S-level)

• Visualisation of the Scooby-domain probability matrix for a sequence can be used to effectively identify regions that are likely to fold into domains or are likely to be unstructured.
Simple protocol for automatic prediction

1. Find highest scoring window (first predicted domain) in probability matrix and sequence region it encapsulates (blue triangle) is removed from the sequence.
2. The resulting sequence fragments are rejoined and the probability matrix recalculated. The smoothing windows that encapsulate the last 15 residues of the N-terminal fragment and the first 15 residues of the C-terminal fragment have their probabilities set to zero (white bands). If the next highest scoring region is found in the red region then the excised domain will be discontinuous, otherwise it will be continuous.

Notes on automatic domain boundary assignment

The Scooby-domain web server (ib.vu.nlprograms/) performs fast, automatic, domain annotation by identifying the most domain-like regions in the query sequence:

1. The highest probability in the domain probability matrix (previous side) represents the first predicted domain.
2. The corresponding stretch of sequence for this domain is removed from the sequence — the first predicted domain will always have a continuous sequence but further domain predictions can encompass discontinuous domains.
3. If the excised domain is at a central position in the sequence, the resulting N- and C-termini fragments are rejoined and the probability matrix recalculated as before. The second highest probability is then found and the corresponding sub-sequence removed.

Evaluating SCOOBYDomain

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (cath)</th>
<th>Specificity (PPV)</th>
<th>Sensitivity (cath)</th>
<th>Specificity (PPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>ScoobyDo</td>
<td>96.5</td>
<td>23.2</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>Domination</td>
<td>56.4</td>
<td>27.9</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>Domaination</td>
<td>56.4</td>
<td>24.0</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td>Larber</td>
<td>67.3</td>
<td>21.8</td>
<td>147.1</td>
</tr>
<tr>
<td></td>
<td>Class</td>
<td>51.2</td>
<td>22.8</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>Best</td>
<td>54.5</td>
<td>23.2</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>ScoobyDo</td>
<td>75.0</td>
<td>63.8</td>
<td>76.7</td>
</tr>
<tr>
<td></td>
<td>Domination</td>
<td>58.8</td>
<td>63.4</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>Domaination</td>
<td>66.4</td>
<td>63.4</td>
<td>87.4</td>
</tr>
<tr>
<td></td>
<td>Larber</td>
<td>79.6</td>
<td>64.2</td>
<td>79.4</td>
</tr>
<tr>
<td></td>
<td>Class</td>
<td>77.0</td>
<td>66.6</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Further assessing SCOOBYDomain predictions

Multilevel smoothing window
- Horizontal axis corresponds to the sequence position
- Vertical axis represents the window length used in the smoothing of sequence hydrophobicity.
- Each position in the matrix corresponds to the average hydrophobicity assigned to the centre of a window during smoothing. (11 amino acid types are considered as hydrophobic: Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp and Tyr)

(c) Each position in the matrix is then converted to a probability that it will fold into a domain, based on the lengths and hydrophobics observed in the distribution of CATH domains.
Wrapping up

- Different approaches to the domain-delineation problem
  - It is a hard problem when having a protein structure at hand
  - It is mind boggling doing it from sequence information alone
- Approaches range from simple window approaches to linker prediction (DomCut) to elaborate consistency-based and 3-D model-reliant prediction (SnapDRAGON)
  - SnapDRAGON is based upon tertiary structure prediction
  - DOMAINATION is based upon (sequence) homology searching
  - SCOOPYDOMAIN is based upon hydrophobicity signals
- Performance still low but results can be very helpful
- Combined iterative methods can improve each of the single methods
- How can these methods be integrated?