Predicting domain features from sequence

Bioinformatics Data Analysis and Tools 2007
Lecture 10
Protein Domain delineation

Content:

• Background
• Linker prediction (DomCut, Elsik)
• Protein domain delineation based on consistency of multiple \textit{ab initio} model tertiary structures (\textit{SnapDRAGON})
• Protein domain delineation based on combining homology searching with domain prediction (\textit{Domaination})
• Domain delineation based on sequence hydrophobicity patterns (\textit{SCOOPY-DOmain})
A domain is a:

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

“Nature is a ‘tinkerer’ and not an inventor” (Jacob, 1977).
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.
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, it 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: aminoimidazole ribonucleotide
Structural domain organisation can be nasty

Pyruvate kinase
*Phosphotransferase*

- β barrel regulatory domain
- α/β barrel catalytic substrate binding domain
- α/β nucleotide binding domain

1 continuous + 2 discontinuous domains

Domain connectivity

Figure 1: Scheme illustrating different types of connectivity in multidomain structures (left) and their sequences (right).

A continuous domain is often an evolutionary module
Domain size

• The size of individual structural domains varies widely
  • from 36 residues in E-selectin to 692 residues in lipoxygenase-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).
Detecting 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
Detecting Structural Domains

• However, approaches encounter problems when faced with highly associated domains (and sometimes also with discontinuous) and many definitions will require manual interpretation.

• Consequently there are discrepancies between assignments made by domain databases (Hadley and Jones, 1999).
Detecting Structural Domains

Early on:

• Interaction of secondary structure: region with weak boundaries are supposed to coincide with domain boundaries (Busetta and Barans, 1984) -- not very successful

• Contact plots: domains are regions with high contact density (Vonderviszt & Simon, 1986) – not very successful
Detecting Structural Domains

More recent methods are better:

- Taylor (1999): will come later during this lecture
Detecting Domains using Sequence only

• Even more difficult than prediction from structure!
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
  - No distinction continuous/discontinuous domain possible
Predicting domain boundaries from linker regions - approaches:

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

\[ S_i = -\ln \left( \frac{f_{i, \text{linker}}}{f_{i, \text{domain}}} \right), \]

where \( i \) denotes the amino acid type and \( f \) the frequencies in either linker or domain.
Predicting domain boundaries from linker regions - approaches:

Bae, Mallick, and Elsik (2005):
- developed a hidden Markov model (HMM) of linker/non-linker sequence regions using a linker index derived from amino acid propensity.
- employed an efficient Bayesian estimation of the model using Markov Chain Monte Carlo (MCMC), particularly Gibbs sampling, to simulate parameters from the posteriors. The model generates a probabilistic output.
- The method was applied to a dataset of protein sequences in which domains and inter-domain linkers had been delineated using the Pfam-A database.
- Prediction results are superior to a simpler method that also uses linker index (DomCut)

\[
\begin{pmatrix}
    p_{00} & p_{01} \\
    p_{10} & p_{11}
\end{pmatrix} =
\begin{pmatrix}
    p_{00} & 1 - p_{00} \\
    1 - p_{11} & p_{11}
\end{pmatrix}
\]

L-L, L-D, D-D, D-L transitions
Integrating protein multiple alignment, secondary and tertiary structure prediction to predict domain boundaries in sequence data

SnapDRAGON

Richard A. George

SnapDRAGON

• Scientific Name
  Antirrhinum majus

  Common Name
  Snapdragon
Protein structure hierarchical levels

**PRIMARY STRUCTURE (amino acid sequence)**

VHLTPEEKSAVTALWGKVNV
EVSQGGRRLVYVPWTQRFF
EFGDLS TDAMGPKVKA
GKVLGSADGALAHLDNLG
ATLSELCFLVDKDPVFRLLG
NVLVCVLAHFGKEFTPVQAA
YQKVVAGVANALAHKYH

**SECONDARY STRUCTURE (helices, strands)**

**QUATERNARY STRUCTURE**

**TERTIARY STRUCTURE (fold)**

Protein structure hierarchical levels

**PRIMARY STRUCTURE** (amino acid sequence)

VHLTPEEKSAVTALWGKVNVD EVGGEALGRLLVVYPWTQRFF ESFGDLSTPDAVMGNPKVKAH GKKVLGAFTPSSDLHLDNLKGT ATLSELHCDKLHVDLENFRLLG NVLVCVLAHHGKEFTPPVQAA YQKVVAGVANALAHKYH

**SECONDARY STRUCTURE** (helices, strands)

**QUATERNARY STRUCTURE**

**TERTIARY STRUCTURE** (fold)
Protein structure hierarchical levels

**PRIMARY STRUCTURE** (amino acid sequence)

VHLTPEEKSAVTALWGVNVD
EVGGEALGRLLVYPWTQRFF
ESFGDLSTPDAVMGNPKVKAH
GKKVLGAFSDGLAHLNDNLKGT
ATLSELHCDKLVDPENFRLLG
NVLVCVLAHFGKEFTPVPQQAA
YQKVVAGVANALAHKYH

**SECONDARY STRUCTURE** (helices, strands)

**TERTIARY STRUCTURE** (fold)

Protein structure hierarchical levels

PRIMARY STRUCTURE (amino acid sequence)

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLALHDNLKGTFTWSELCHELCDKHLVDPEFRLLGNVLVCLAHFGKEFTPPVQAYQKVVAVANALAHKYH

SECONDARY STRUCTURE (helices, strands)

QUATERNARY STRUCTURE

TERTIARY STRUCTURE (fold)
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 for 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

SnapDragon

Multiple alignment

Predicted secondary structure
\textit{CCHHHCCCEE}

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

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)

<table>
<thead>
<tr>
<th>Consistency</th>
<th>5</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>3</th>
<th>7</th>
<th>4</th>
<th>3</th>
<th>3</th>
<th>8</th>
<th>55</th>
<th>8</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1n97a</td>
<td>yfp</td>
<td>g</td>
<td>u</td>
<td>q</td>
<td>r</td>
<td>i</td>
<td>c</td>
<td>g</td>
<td>r</td>
<td>d</td>
<td>f</td>
<td>a</td>
<td>l</td>
</tr>
<tr>
<td>1gwia</td>
<td>histghmphvcpgaalismea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1jipa</td>
<td>histgqpithcmgrplakleeg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1fba</td>
<td>p</td>
<td>l</td>
<td>g</td>
<td>f</td>
<td>u</td>
<td>d</td>
<td>h</td>
<td>r</td>
<td>c</td>
<td>i</td>
<td>a</td>
<td>e</td>
<td>h</td>
</tr>
<tr>
<td>1cpt</td>
<td>hlgf</td>
<td>gwa</td>
<td>ah</td>
<td>m</td>
<td>c</td>
<td>g</td>
<td>q</td>
<td>h</td>
<td>l</td>
<td>a</td>
<td>k</td>
<td>l</td>
<td>e</td>
</tr>
<tr>
<td>1n40a</td>
<td>hlafgrcqhtcpgsalignrha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11ka</td>
<td>h</td>
<td>v</td>
<td>a</td>
<td>r</td>
<td>g</td>
<td>h</td>
<td>u</td>
<td>v</td>
<td>h</td>
<td>c</td>
<td>g</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>1qmqa</td>
<td>h</td>
<td>t</td>
<td>t</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>c</td>
<td>g</td>
<td>q</td>
<td>h</td>
<td>l</td>
<td>a</td>
<td>r</td>
</tr>
<tr>
<td>1io7a</td>
<td>hlsgspithcligaplarraye</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1jpza</td>
<td>fkpfngnqra</td>
<td>c</td>
<td>g</td>
<td>q</td>
<td>q</td>
<td>f</td>
<td>a</td>
<td>l</td>
<td>h</td>
<td>e</td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1e9xa</td>
<td>wipfgasvhrncvga</td>
<td>a</td>
<td>f</td>
<td>a</td>
<td>i</td>
<td>m</td>
<td>q</td>
<td>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1dt6a</td>
<td>fmpfsankrmcvgeglarre</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SNAPDRAGON
Domain boundary prediction protocol using sequence information alone (Richard George)

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

[31] 21 May 2007
• The Cα distance matrix is divided into smaller clusters.

• Separately, each cluster is embedded into a local centroid.

• The final predicted structure is generated from full embedding of the multiple centroids and their corresponding local structures.
Lysozyme 4lzm

PDB

DRAGON
Phosphatase 2hhm-A

PDB

DRAGON

Taylor method (1999)  
DOMAINT-3D

3. Assign domain boundaries to each of the 3D models (Taylor, 1999)
   • 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 neighbourhood 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

Taylor method (1999)

repeat until convergence

if $41 < (5+6+56+78+89)/5$
then Res $41\leftarrow 42$ (up 1)
else Res $41\leftarrow 40$ (down 1)
Taylor method (1999)

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

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

‘Res number’

Sequence location

continuous

discontinuous

[38] 21 May 2007
SNAPDRAGON
Domain boundary prediction protocol using sequence information alone (Richard George)

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 = \( \sum_{1 \leq i \leq l} S_i \times W_i \)

Where \( W_i = (p - |p-i|)/p^2 \) and \( p = \frac{1}{2}(n+1) \).
It follows that \( \sum_l W_i = 1 \)

SNAPDRAGON

Statistical significance:
- Convert peak scores to Z-scores using $z = (x\text{-mean})/\text{stdev}$
- If $z > 2$ then assign domain boundary

Can further test statistical significance using random models:
- Test hydrophobic collapse given distribution of hydrophobicity over sequence
- Make 5 scrambled multiple alignments (MSAs) and predict their secondary structure
- Make 100 models for each MSA
- Compile mean and stdev from the boundary distribution over the 500 random models
- If observed peak $z > 2.0$ stdev (from random models) then assign domain boundary
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 ±10 residues from true boundary)
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 taken and sequence split equally as in baseline method I
Average prediction results per protein

<table>
<thead>
<tr>
<th></th>
<th>Continuous set</th>
<th>Discontinuous set</th>
<th>Full set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SnapDRAGON</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td>63.9 (± 43.0)</td>
<td>35.4 (± 25.0)</td>
<td>51.8 (± 39.1)</td>
</tr>
<tr>
<td>Success</td>
<td>46.8 (± 36.4)</td>
<td>44.4 (± 33.9)</td>
<td>45.8 (± 35.4)</td>
</tr>
<tr>
<td><strong>Baseline 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td>43.6 (± 45.3)</td>
<td>20.5 (± 27.1)</td>
<td>34.7 (± 40.8)</td>
</tr>
<tr>
<td>Success</td>
<td>34.3 (± 39.6)</td>
<td>22.2 (± 29.5)</td>
<td>29.6 (± 36.6)</td>
</tr>
<tr>
<td><strong>Baseline 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td>45.3 (± 46.9)</td>
<td>22.7 (± 27.3)</td>
<td>35.7 (± 41.3)</td>
</tr>
<tr>
<td>Success</td>
<td>37.1 (± 42.0)</td>
<td>23.1 (± 29.6)</td>
<td>31.2 (± 37.9)</td>
</tr>
</tbody>
</table>

Coverage is the % linkers predicted (TP/TP+FN)
Success is the % of correct predictions made (TP/TP+FP)
Average prediction results per protein
SnapDRAGON

• Uses consistency in the absence of standard of truth
• 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
• Strategy is now used by the Baker group (UW, Seattle)
Integrating protein sequence database searching and on-the-fly domain recognition

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).
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
A Profile Matrix (Position Specific Scoring Matrix - PSSM)

This is the same as a profile without position-specific gap penalties.
PSI BLAST:
Constructing the Profile Matrix

Figure from: Altschul et al. Nucleic Acids Research 25, 1997
PSI-BLAST steps in words

- 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.
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

Paste your query sequence

Switch this off for default run
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.
‘X’ residues denote low-complexity sequence fragments that are ignored.
Sequence searching

QUERY

DATABASE

POSITIVES

NEGATIVES

True Positive

False Positive

True Negative

False Negative

True Positive

True Negative

[58] 21 May 2007
 PSI-BLAST

Strategy: Combine C- and N-termini of local alignments to delineate domain boundaries
Submit a query sequence

PSI-BLAST database search

Filter low complexity regions in database sequences with SEG

Delineate domains

Exit

No domains

Filter homologous sequences using OBSTRUCT and create multiple alignments using PRALINE

Chop and Join Domains
Post-processing low complexity
Remove local fragments with > 15% LC
Identifying domain boundaries

Sum N- and C-termini of gapped local alignments

True N- and C-termini are counted twice (within 10 residues)

Boundaries are smoothed using two windows (15 residues long)

Combine scores using biased protocol:

\[
\begin{align*}
\text{if} & \quad Ni \times Ci = 0 \\
\text{then} & \quad Si = Ni + Ci \\
\text{else} & \quad Si = Ni + Ci + (Ni \times Ci)/(Ni + Ci)
\end{align*}
\]
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. (remove N- and C- termini)
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.
Identifying continuous and discontinuous domains

• Each segment is assigned an independence score (In).
  If In > 10% the segment is assigned as a continuous domain.
• An association score is calculated between non-adjacent fragments by assessing the shared sequence hits to the segments. If score > 50% then segments are considered as discontinuous domains and joined.
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 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.
Domain boundary prediction accuracy

- 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% (±44.6%) correct boundary predictions per protein
Domain boundary prediction accuracy

- 23.3% of all linkers found in 452 multidomain proteins. *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
- 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.
Two reference sets based on individual domain searches (using known domains)

- Reference set 1: consists of database sequences for which PSI-BLAST finds all domains contained in the corresponding full length query.
- Reference set 2: consists of database sequences found by searching with one or more of the domain sequences
- Therefore set 2 contains many more sequences than set 1
Sequences found over Reference sets 1 and 2

<table>
<thead>
<tr>
<th></th>
<th>PSI-BLAST vs Ref set 1</th>
<th>DOMAINATION vs Ref set 1</th>
<th>PSI-BLAST vs Ref set 2</th>
<th>DOMAINATION vs Ref set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq's found</td>
<td>28581</td>
<td>28921</td>
<td>67300</td>
<td>73274</td>
</tr>
<tr>
<td>Seq's missed</td>
<td>618</td>
<td>278</td>
<td>13542</td>
<td>7568</td>
</tr>
<tr>
<td>% missed</td>
<td>2.12</td>
<td>0.95</td>
<td>16.8</td>
<td>9.36</td>
</tr>
</tbody>
</table>

Note that PSI-BLAST and DOMAINATION were run over full sequences in Ref sets 1 and 2
Reference 1

• PSI-BLAST finds 97.9% of sequences
• Domaination finds 99.1% of sequences

Reference 2

• PSI-BLAST finds 83.2% of sequences
• Domaination finds 90.6% of sequences
SSEARCH significance test

• Verify the statistical significance of database sequences found by relating them 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.

• This filter will lose distant homologies (bad E-values).

• 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.
Scooby-domain: prediction of globular domains in protein sequence

Richard A. George\textsuperscript{1,2}, Kuang Lin\textsuperscript{3} and *Jaap Heringa\textsuperscript{4}

\textsuperscript{1} Inpharmatica Ltd, 60 Charlotte Street, London W1T 2NU UK
\textsuperscript{2} European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
\textsuperscript{3} Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill NW7 1AA, UK
\textsuperscript{4} Centre for Integrative Bioinformatics (IBIVU), Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
* 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.
• First plot: the number of CATH domains as a function of their hydrophobicity and domain length.

• Second plot: the average CATH domain hydrophobicity minus the average hydrophobicity for randomised sequences (generated from a random selection of residues from sequences in the CATH database).

• Information is used to create partition density function for domain likelihood
CATH domains Randomized domain sequences

CATH domains minus Randomized domain sequences
(b) 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 hydrophobicities observed in the distribution of CATH domains.
(d) i. The highest scoring window (first predicted domain) is identified in the probability matrix and the sequence region it encapsulates (blue triangle) is removed from the sequence. ii. The resulting sequence fragments are rejoined and the probability matrix recalculated. iii. 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.
Automatic domain boundary assignment

• The **Scooby-domain** web server (ibi.vu.nl/programs/) performs fast, automatic, domain annotation by identifying the most domain-like regions in the query sequence:
  • The highest probability in the domain probability matrix represents the first predicted domain.
  • The corresponding stretch of sequence for this domain is removed from the sequence -- the first predicted domain will always have a continuous sequence and further domain predictions can encompass discontinuous domains.
  • 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.
<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Accuracy (PPV)</th>
<th>Sensitivity</th>
<th>Accuracy (PPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScoobyDo</td>
<td>50.5</td>
<td>23.2</td>
<td>51.8</td>
<td>30.8</td>
</tr>
<tr>
<td>Domaination</td>
<td>59.6</td>
<td>27.6</td>
<td>59.8</td>
<td>29.5</td>
</tr>
<tr>
<td>Linker</td>
<td>42.7</td>
<td>14.8</td>
<td>42.7</td>
<td>14.8</td>
</tr>
<tr>
<td>Class</td>
<td>41.6</td>
<td>22.9</td>
<td>40.1</td>
<td>25.1</td>
</tr>
<tr>
<td>Best</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScoobyDo</td>
<td>75.1</td>
<td>44.4</td>
<td>76.7</td>
<td>50.1</td>
</tr>
<tr>
<td>Domaination</td>
<td>88.8</td>
<td>44.4</td>
<td>87.4</td>
<td>47.4</td>
</tr>
<tr>
<td>Linker</td>
<td>79.4</td>
<td>34.1</td>
<td>79.4</td>
<td>34.1</td>
</tr>
<tr>
<td>Class</td>
<td>71.0</td>
<td>46.6</td>
<td>70.9</td>
<td>48.0</td>
</tr>
</tbody>
</table>

Two measures are used to score predictions: percentage of real boundaries predicted (sensitivity) and percentage of correct predictions made (accuracy). ‘N- and C-termini weighting’ are predictions made with increased probability of domain boundaries at the ends of the protein sequences. ‘Domaination’ are results for ScoobyDo predictions made with added information from Domaination. ‘Linker’ are results for ScoobyDo predictions made with added information from the interdomain linker propensities from the Linker database. ‘Class’ are ScoobyDo predictions made using three smoothing windows to separately predict all-α, all-β and α-β domains. ‘First’ is the highest probability prediction made. ‘Best’ is the best of ten predictions made.
Enter a protein sequence here:

TTPQEDGFLRLKIASKEKIARDIWSFELTDPCAPLPFFEAGANLTVWVVPNGSFRTVSLNDSQLRNVYVIAVFRDSTNGGRGSISFDTSCDAVEVSLPRNEFPILDKRAKSFLVACIGITPMISMARQLFAEGLSRSLYLTRDPEGTFDFDELTSDEWSDSVKIHHDGDPTKFDWFSVFEKSKPAQHVVCCGPQALMDTVDMMTHWPSGTVHFESFGATNTNAERTNPFTRLSRSFSFIPANRSILEVLRDANVRVPSCSSGETCSCKTACRCSGATMGHARDMVLRDDKGTQIMVCVSRKSAELVLDL

res = 157  win = 112  value = 0.870  limit = 102-212

21 May 2007
Identifying foldable regions in protein sequence from the hydrophobic signal

Chi N. I. Pang\textsuperscript{a}, Kuang Lin\textsuperscript{b}, Merridee A. Wouters\textsuperscript{a,c}, Jaap Heringa\textsuperscript{d}, Richard A. George\textsuperscript{a}

\textsuperscript{a}Structural & Computational Biology Program, Victor Chang Cardiac Research Institute, Sydney, NSW, Australia \textsuperscript{b}Biomathematics and Statistics, Scotland, JCBM, The King’s Building, Edinburgh, EH9, 3JZ, Scotland, UK \textsuperscript{c}School of Biotechnology & Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia \textsuperscript{d}Centre for Integrative Bioinformatics, Faculty of Sciences and Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

Improvements:

- Use Multiple Sequence Alignments and average prediction results
- Use A* combining domain delineation protocol for 10 top-predictions
The Scooby+MSA prediction for the hyperthermostable D-ribose-5-phosphate isomerase from *Pyrococcus horikoshii* (PDB 1LK5, chain A). a) The structure of 1LK5, coloured according to the linker prediction by Scooby-Domain. The corresponding predictions are 136 and 207. The CATH domain annotation shows that it consists of two domains, a discontinuous domain made of two segments 1-128 (green) and 208-229 (blue); and the continuous domain 129-207 (red). b) The Scooby-Domain plot for 1LK5.
Scooby-domain prediction

![Graphs showing Scooby-domain prediction](image)
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)

• Performance still low but results can be very helpful

• Domaination: combined iterative methods can improve each of the single methods