Genome scale Comparison of ligand binding sites in protein structures: Algorithms and applications in drug discovery

Nagasuma Chandra
Indian Institute of Science
Bangalore, India
Nagasuma Chandra, Research Overview

Pathway Modelling
Reactome Modelling
Protein-Protein influences
Interactome Modelling

Algorithms for Structure Analysis

SYSTEMS BIOLOGY

STRUCTURAL BIOINFORMATICS

Drug Discovery

Drug Target Identification
Druggability Assessment
Lead Identification

Drug Resistance
Host-Pathogen Modelling
Modelling drug effects
Drug failure Analysis
Modeling Metabolism in *M. tuberculosis*

<table>
<thead>
<tr>
<th>Genes</th>
<th>661</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>543</td>
</tr>
<tr>
<td>Reactions (Intra Systems)</td>
<td>939</td>
</tr>
<tr>
<td>Reactions (Exchange)</td>
<td>88</td>
</tr>
<tr>
<td>Gene Association Reaction</td>
<td>77%</td>
</tr>
<tr>
<td>Metabolites</td>
<td>828</td>
</tr>
<tr>
<td>Average Confidence Level</td>
<td>2.31</td>
</tr>
</tbody>
</table>

**Growing bacteria *In silico* under different media**

**In silico gene deletions**

**Essential nutrients required for growth**

**Response to other Perturbations-Nutrient Uptake**

**Insights obtained**

- Gene Essentiality: 220 essential genes
- Nutrient essentiality:
- Hard coupled reaction sets: groups of reactions that are forced to operate in unison due to mass conservation and connectivity constraints)
- Fatty Acid Metabolism & Lysine Metabolism

Raman *et al.*, 2008
Drug Resistance Pathways
Abstraction of the flow of information that leads to Drug Resistance in TB bacilli

- Network of shortest paths from MAP to Resistance Genes
- 616 nodes and 1,683 edges
- Paths scored based on edge frequency, up-regulation of source and target nodes

Raman and Chandra, 2008, BMC Microbiology
A Boolean model of HPIs developed, Simulations to capture a variety of scenarios

Raman, Bhat & Chandra, Mol. Biosyst, 2010
Modelling iron homeostasis in M. tuberculosis
Soma Ghosh, KVS Prasad, Sarswathi Vishveshwara, Nagasuma Chandra

Integration of Systems Perspective with Structural level detail

Raman, Kalidas, Chandra, 2008
Comparison of protein molecules

How to compare?
- Annotation - keyword
- Function identification
  - Molecular level
  - Cellular level
- Sequence - Alignment
- Structure – Fold comparison
Problems with these approaches

- Dissimilar Sequences - Similar Function → benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC)
- Similar sequences - Different Function → Steroid delta isomerase; nuclear transporter 2; scytalone dehydratase

**Similar Structures - Different Function** → triose phosphate isomerase and FMN-linked oxidoreductases

**Dissimilar Structures – Similar Function** → ATP binding proteins from different SCOp families; C-type lectin and bulb lectin

Chymotrypsin & Subtilisin
What really matters for a protein molecule is its function and not what means it uses to achieve it!
It’s the **meaning** that counts....

**Whether two proteins can recognize the same molecules**
Molecular recognition
<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession</th>
<th>Sequence Description</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GARL 1</td>
<td>39</td>
<td>YAGQSLDVPEYHLIM</td>
<td>EDGNLVLDHSTAVTWSNTDIPG</td>
</tr>
<tr>
<td>GARL 2</td>
<td>71</td>
<td>ASHTDPGBKGGCKAVLSDGNFVVDAEGRSLWASHSVRG</td>
<td>--</td>
</tr>
<tr>
<td>GARL 3</td>
<td>103</td>
<td>ASHSTVRCGNYLVLMDDCNYLVDV-KFIVATNGSL</td>
<td>--</td>
</tr>
<tr>
<td>SNOW 1</td>
<td>34</td>
<td>STGEFLNYGSFVIMDDCNYLVDDVDKFIWATNGSL</td>
<td>--</td>
</tr>
<tr>
<td>SNOW 2</td>
<td>65</td>
<td>ATNTRGLS-RSCFLSMDDCNYLVNPNSKNPFWASNTGQ</td>
<td>--</td>
</tr>
<tr>
<td>SNOW 3</td>
<td>97</td>
<td>ASHTGVC1N-GNYVICTDGNLVDDNKSIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>HIPE 1</td>
<td>111</td>
<td>ASHPEVL1N-GNYVICTDGNLVDDNKSIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>HIPE 2</td>
<td>18</td>
<td>ASNTGLA-RGCGLSMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>HIPE 3</td>
<td>50</td>
<td>ASNTSGEN1N-GNYVICTDGNLVDDNKSIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>NARC 1</td>
<td>35</td>
<td>STGQFLSYGSYVFIMDDCNYLVDV-KFIVATNGSL</td>
<td>--</td>
</tr>
<tr>
<td>NARC 2</td>
<td>66</td>
<td>ATNTRGLS-RSCFLSMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>NARC 3</td>
<td>98</td>
<td>ASHTGVC1N-GNYVICTDGNLVDDNKSIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>ONIO 1</td>
<td>26</td>
<td>YAGQSLVYEYQTFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>ONIO 2</td>
<td>58</td>
<td>SNTGVTGKGNCIRA-VADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>ONIO 3</td>
<td>90</td>
<td>ASNSRGRG-NGYILVDV-KNVRGIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>RAMS 1</td>
<td>71</td>
<td>SNTGVR3GRNCIRA-VADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>RAMS 2</td>
<td>103</td>
<td>ASQSSRGN1-GNYILVDDVCNVRGIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>URSI 1</td>
<td>38</td>
<td>YAGQSELLPYKLIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>URSI 2</td>
<td>70</td>
<td>TNGSVELGRNCIRA-VADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>URSI 3</td>
<td>102</td>
<td>ASNSIKGN1-GNYILVDV-KNVRGIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>LEAK 1</td>
<td>41</td>
<td>YAGQSDLVPEYHLIM</td>
<td>EDGNLVLDHSTAVTWSNTDIPG</td>
</tr>
<tr>
<td>LEAK 2</td>
<td>67</td>
<td>ATNTRGLS-RSCFLSMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>LEAK 3</td>
<td>105</td>
<td>ATNTRGLS-RSCFLSMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>SHAL 1</td>
<td>37</td>
<td>YAGQSLVYEYQTFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>SHAL 2</td>
<td>69</td>
<td>SNTGVTGKGNCIRA-VADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>SHAL 3</td>
<td>101</td>
<td>ASNSRGRG-NGYILVDV-KNVRGIAWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>TULI 1</td>
<td>54</td>
<td>YAGQSLTVEYQTFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>TULI 2</td>
<td>85</td>
<td>ASGSNDLG-SCGYVLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>TULI 3</td>
<td>117</td>
<td>ASQTHQAE-GNYVLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CLIV 1</td>
<td>38</td>
<td>SFGSELSHGRVFMDDCNYLVDV-KFIVATNGSL</td>
<td>--</td>
</tr>
<tr>
<td>CLIV 2</td>
<td>69</td>
<td>ATNTRGLS-RSCFLSMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CLIV 3</td>
<td>102</td>
<td>SNTGGERANYVLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>EPIN 1</td>
<td>33</td>
<td>GTGGSLAAGGYIFKADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>EPIN 2</td>
<td>70</td>
<td>ASHTGNC1N-GNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>EPIN 3</td>
<td>101</td>
<td>ASNTQLTDGNNFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>LIST 1</td>
<td>38</td>
<td>NTGSQSLTDDGNNFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>LIST 2</td>
<td>69</td>
<td>ASQTHQAE-GNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>LIST 3</td>
<td>102</td>
<td>SNTGVEFLYVTLIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CMB 1</td>
<td>40</td>
<td>NPGQSLTSNGVDLMADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CMB 2</td>
<td>71</td>
<td>SSQTYGS-SCGYVLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CMB 3</td>
<td>103</td>
<td>ASNTNGL1N-GNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CUCK 1</td>
<td>37</td>
<td>NTDGRLNGETLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>CUCK 2</td>
<td>157</td>
<td>YGQSLTSNGVDLMADDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>POLY 1</td>
<td>40</td>
<td>FSQHSCTGSYRLIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>POLY 2</td>
<td>103</td>
<td>QNTNBEKE-DHYLMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
<tr>
<td>ALOE 1</td>
<td>11</td>
<td>HENQYISGYFIMDDCNYLVIDPNSNPFWASNTDGE</td>
<td>--</td>
</tr>
</tbody>
</table>

Multiple alignment of bulb lectins indicating the mannose binding sequence motif
A structural motif specific for mannose recognition
ADP-AlF₄ In The Binding Cleft of MtRecA
Structural Bioinformatics

- PDB
- Family-specific databases (SCOP, CATH, CAI-H)
- Structural Bioinformatics
  - Fold Comparison
  - Binding site detection
  - Structural motifs
  - Substructures
  - Interaction profiles
- Protein-Ligand Docking
  - Protein-Ligand Pharmacophores
  - 3D and 4D Pharmacophores
- Lead Identification and Optimization
Structural Bioinformatics: *Development of 5 novel algorithms integrated into PocketSuite*
Binding site Prediction Methods:

Homology-based methods
- Alignment with known sites
- Conservation

Sequence-based methods
- Motifs

Structure-based methods
- Geometric Chemical
- Hybrid methods

Machine learning
Binding site prediction Methods

- Agglomerative clustering of atoms having greater depth from the alpha triangles
- Fitting spheres between neighbouring atoms and clustering spheres based on size
- Filling surface with layer of probes (PASS, BIOSUITE)
- Determining convex hull (Alpha spheres, CASTp, MOE)
- Computing interaction energy of probe at various positions around protein + clustering

Non-grid

Indexing subspaces

Grid

Existing methods

Force field

Molecular dynamics based

Pattern matching

Clustering of grid cells having similar force field potential

- Scanning grid along multiple directions (LIGSITE)
- Visibility criteria

P.E. - energy
- Goodfors's method
- PocketFinder
- Q-SiteFinder

Data mining and machine learning
- Structural and sequence motifs
<table>
<thead>
<tr>
<th>Method</th>
<th>Algorithm</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>POCKET</td>
<td>Scan the grid along 3 dimensions</td>
<td>(Levitt and Banaszak, 1992)</td>
</tr>
<tr>
<td>Ligsite</td>
<td>Scanning grid along 3 axes and 4 diagonals</td>
<td>(Hendlich et al., 1997)</td>
</tr>
<tr>
<td>LigSiteCSC</td>
<td>Similar to Ligsite but with residue conservation information for each set of residues to occur in site</td>
<td>(Huang and Schroeder, 2006)</td>
</tr>
<tr>
<td>LigandFit</td>
<td>Eraser to swipe the grid cells to demarcate cells belonging to a grove</td>
<td>(Venkatachalam et al., 2003)</td>
</tr>
<tr>
<td>PASS</td>
<td>Filling up surface of protein by multiple layers of probes and retaining probes with high burial count</td>
<td>(Brady and Stouten, 2000)</td>
</tr>
<tr>
<td>CASTp</td>
<td>Fill the interatomic regions by spheres and cluster moderately sized spheres</td>
<td>(Liang et al., 1998)</td>
</tr>
<tr>
<td>VOIDOO</td>
<td>Similar to VOIDOO</td>
<td>(Kleywegt and Jones, 1994)</td>
</tr>
<tr>
<td>SURFNET</td>
<td>Determine depressions on the surface of protein by placing spheres between pairs of atoms.</td>
<td>(Glaser et al., 2006)</td>
</tr>
<tr>
<td>APROPOS</td>
<td>Find clusters of atoms with depth from surface of protein</td>
<td>(Peters et al., 1996)</td>
</tr>
<tr>
<td>Goodford’s method</td>
<td>Clustering of grid cells with higher energy values</td>
<td>(Goodford, 1985)</td>
</tr>
<tr>
<td>Q-SiteFinder</td>
<td></td>
<td>(Laurie and Jackson, 2005)</td>
</tr>
<tr>
<td>PocketFinder</td>
<td></td>
<td>(Jianghong et al., 2005)</td>
</tr>
</tbody>
</table>
Geometry based

a. POCKET, LIGSITE, LIGSITE
c

b. SURFNET

---

c. CAST

d. PASS

---

Alpha shape

Triangles flow

Dcryst 2007Sep23
PocketDepth
Grid based binding site prediction method

Figure 1

(a)

(b)

(c)
Grid Bar Generation

Fill inter-atomic (surface atoms) regions with grid bars. A grid bar \( \{x, y\} \in A : GB(x, y) \subset G \) between pair of atoms \( x, y \). A grid bar is valid only if does not intersect an atom. Obtain set of all valid grid bars

\[ \{(\forall a, b \in S)(\exists c \in (A - S) : cell(c) \in GB(a, b))\} \]
Rendering of DepthFactor as temperature

Update traversal counter, called Depth Factor, of each grid cell in a valid $GB(a, b)$
$(\forall c \in GB(a, b))c\text{.depth} \leftarrow c\text{.depth} + 1$

Cluster grid cells based on Depth Factor and spatial proximity (DBSCAN)

Partition the whole of the set of grid cells $G$ into non-overlapping clusters

$S^C = C_1,...,C_n : C_i \cap C_j = \emptyset$ where $S^C$ denotes a set of clusters 1...n
CLUSTERING METHODS

Partitioning Clustering
- Distance-based
  - K-means
  - CLARANS
- Model-based
  - EM
- Density-based
  - DBSCAN
  - CLIQUE

Hierarchical Clustering
- Distance-based
- Density-based
  - Neighborhood-based
  - Single-link
  - Graph-based
  - OPTICS (forthcoming)
  - Grid-based
**DBSCAN clustering**

FUNCTION DBSCAN(point \(p, c, N\))

\(p\) is a point and \(c\) is cluster number

**if** \(|S = q : d(q, p) \leq d_{\text{threshold}} \land \epsilon| \geq N** then

\(p_c \leftarrow c\)

**call** DBSCAN\((q) : (q \in S)\)

**end if**
Clustering based on Depth Factor

Each grid cell in a cluster \( C_i (\forall i \in [1...n]) \) satisfies the depth and density requirements
\[
(\forall (c \in C_i) : |\{(\forall c' \in C_i) | distance(c, c') \leq \rho \}| \geq N \land DF(c) \geq \Delta
\]
where \( \land \) denotes logical AND
where \( \rho, N \) are radius and number of points within radius (DBSCAN parameters); and \( DF(c) \) is the Depth Factor and \( \Delta \) is the imposed threshold
In 82% and 94% of proteins (*PDBBind* 1091) top 5 and 10 ranked clusters overlapped with crystal ligand

In dimers ranks 1 and 2 corresponded to ligand locations

Binding Site Prediction Algorithms: PocketDepth Performance

<table>
<thead>
<tr>
<th>PDB</th>
<th>CASTp</th>
<th>PocketDepth</th>
<th>LigandFit</th>
<th>LigSite&lt;sup&gt;CSC&lt;/sup&gt;</th>
<th>QSiteFinder</th>
</tr>
</thead>
<tbody>
<tr>
<td>1q55</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>1,3,12</td>
<td>1</td>
<td>One</td>
<td>One of top 10</td>
<td>1.2</td>
</tr>
<tr>
<td>1a72</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>2,11,13,17</td>
<td>1.3</td>
<td></td>
<td>2.7,10</td>
<td></td>
</tr>
<tr>
<td>1ais</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>1,2,6,7,12,14,16,21,23</td>
<td>1.3,6,7,15,19,20</td>
<td>2.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1s1m</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>2,3,13,14,28</td>
<td>2.3,5</td>
<td>1.2,6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2pel</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>2,5,13,16</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2g88</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>ranks</td>
<td>1,2,3,4,7,30,48,50</td>
<td>1.2,5</td>
<td></td>
<td>2.5,7</td>
<td></td>
</tr>
</tbody>
</table>

Kalidas & Chandra, 2008
JSB
Binding Site Comparison
Need..

Binding site comparison can

- Predict Important residues in a protein binding site
- Predict the function of a hypothetical protein.
- Predict the similarity between proteins.
Context of Binding Comparison...

- Sequence or structural similarity && Not same molecular function
- Same function && No fold similarity


- Necessity of binding site comparison methods
  - Understanding protein function
  - Understanding side effects of drugs
Challenges in site comparison

- **Point set superposition**
- **Binding site** → **Set of points (atoms/residues)**
- **Determining point-point correspondences**
- **Topology Undefined; Size of ‘match’ small & unknown (‘Indels’ possible)**
- **Involves costly least squares evaluation of rotation & translation matrices**
- **Many possible correspondences**

**Geometric Hashing; Maximal Common Sub-graph Search; Depth First Traversal** (incrementally determine correspondences)
Comparison of a pair of binding sites involves three aspects:

(a) representation of each site as sorted lists of distances between chosen points,

(b) alignment of two sets of distance lists and

(c) choosing a scoring scheme for arriving at a final score
Description of the site

• Global features
  – Volume
  – Surface area
  – Number of polar/non-polar atoms/residues

• Shape Descriptors: Frame-invariant representations
  – Image moments
  – Spherical harmonics
  – All pair sorted distance sequences \textit{(PocketMatch)}

(Morris et al., 2005, Bioinformatics; Gold and Jackson, 2006, NAR; CavBase - Kuhn et al., 2007, CHEMMEDCHEM; PINTS – Stark et al., 2003, NAR; SPASM & RIGOR – Gerard et al., 1999, JMB; Binkowski et al., 2003, JMB; Morris et al., 2005, Bioinformatics; Nagano et al., 2002, JMB; Kunin et al., 2001, JMB; Campbell et al., 2003, An et al., 2005)
Tools for Binding site comparison:

- **PocketMatch** - A new algorithm to compare binding sites in protein structures
- **CavBase**
- **SitesBase** - a database for structure-based protein–ligand binding site comparisons
- **CPASS** - Comparison of Protein Active-Site Structures
- **PINTS** - Patterns in Non-homologous Tertiary Structures
- **Spasm/RIGOR**
- **SMAP-WS Pairwise Comparison /SMAP-WS Database Search**
- **SiteSorter** - N-by-N Binding Site Similarity Assessment
- **SLiC** - Site-Ligand Contact Analysis and Binding Mode Similarity Assessment
- **MAPPIS (Multiple Alignment of Protein-Protein InterfaceS (PPIs))** - Recognizes spatially conserved chemical interactions shared by a set of PPIs
- **MULTIBIND (Multiple Alignment of Protein Binding Sites)** - Recognizes Spatial Chemical Binding Patterns Common to a Set of Protein Structure
PocketMatch Algorithm

Number of matching distance elements

\[ \text{PMScore} = \frac{\sum_{i=1}^{90} \text{Count}_i}{\max(|S_1|, |S_2|)} \]
PocketMatch Algorithm
http://proline.physics.iisc.ernet.in/pocketmatch/

3 types of points (CA,CB,CNTR)
5 types of residue groups (AVILGP; KRH; DE; YFW; CSTQN)

\( (3*(3-1)/2+3)*(5*(5-1)/2+5) \rightarrow 90 \) lists
120 lists are possible and yielded similar results.

PocketMatch: A new algorithm to compare binding sites in protein structures
Kalidas Yeturu and Nagasuma Chandra, 2008, BMC Bioinformatics
PocketMatch implementation

- Sites extracted → around (4Å) ligand/predicted pocket
- Complete residues → representative points → Sorted Distance lists
- MPI version (C language)
- Run on IBM Bluegene utilizing 1024 processors
Perturbation studies

Validation with respect to random perturbation of positions of site-points

Random perturbations of site points for (a) ligand (PP8) with 54 PMScores for perturbed sites with respect to its original site for different extents of perturbations (RMSD) are shown at different values of (1.0-green, 0.5-red, 0.25-cyan, 0.125-blue, 0.01-yellow)

Type perturbation
Superposition of sites with (a) High PMScores (80.9% for 1H8H-ATP and 1W0K-ADP) and (b) low PMScores (25.8% for 1H8H-ATP and 1H8H-ADP)
Validation

ATP-ADP Similar and dissimilar sites

SCOP VS PM

Known similarities in tetramers
Detection of part-similarities by PocketMatch.
Examples illustrating binding of different ligands in essentially the same binding pocket, but with different orientations. The part-similarities in these were identified correctly by PocketMatch. Binding of different trypsin inhibitors (stick models) complexed to trypsin variants (wire) as in PDB entries (a) 1GJC and 1V2Q and (b) 1GJC and 2AYW.
Need for site alignment

• Different folds exhibit similar binding sites
  – Ex – cofactor binding sites HEM, NAD, FAD
• Difficult to detect local similarities by human – error prone
• Structural motifs – determining function
Challenges in alignment

• Many possible local similarities exist
  – Exhaustive enumeration is impractical

• Finding out the best is tough
  – Quantify when does an expert call superposition ‘good’

• What level to consider structural match
  – Atomic, Residue, C-alpha
Superpositions of FAD binding sites 1C0P and 1HYU
POCKETALIGN
PocketAlign

Alignment of binding sites

Yeturu and Chandra, JCIM (ACS), 2011, In press
Schematic

SITE 1

1. Extract co-ordinates of site residues

<table>
<thead>
<tr>
<th>A</th>
<th>LYS</th>
<th>ASN</th>
<th>21.04</th>
<th>15.73</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.83</td>
<td>0.83</td>
<td>11.56</td>
<td>13.50</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>0.13</td>
<td>11.04</td>
<td>11.29</td>
</tr>
<tr>
<td>4</td>
<td>1.09</td>
<td>22.22</td>
<td>11.83</td>
<td></td>
</tr>
</tbody>
</table>

SITE 2

1. Compute $C_{\alpha}; C_{\alpha}$ distance matrix

<table>
<thead>
<tr>
<th>A</th>
<th>LYS</th>
<th>ASN</th>
<th>18.34</th>
<th>11.96</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.84</td>
<td>0.84</td>
<td>15.70</td>
<td>13.73</td>
</tr>
<tr>
<td>3</td>
<td>10.70</td>
<td>0.0</td>
<td>3.78</td>
<td></td>
</tr>
</tbody>
</table>

2. Generate Sorted List of Distances

<table>
<thead>
<tr>
<th>A</th>
<th>0.0</th>
<th>4.24</th>
<th>8.34</th>
<th>14.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
<td></td>
</tr>
</tbody>
</table>

3. Align Sorted Lists & Generate Geometric Perspective Scoring (GPS) matrix for all alignments

<table>
<thead>
<tr>
<th>A</th>
<th>0.0</th>
<th>4.24</th>
<th>8.34</th>
<th>14.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
<td></td>
</tr>
</tbody>
</table>

Alignment Score = 2

<table>
<thead>
<tr>
<th>A</th>
<th>0.0</th>
<th>8.84</th>
<th>10.78</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
</tr>
</tbody>
</table>

Alignment Score = 3

4. Construct the GPS x BL matrix

<table>
<thead>
<tr>
<th>A</th>
<th>0.0</th>
<th>4.24</th>
<th>8.34</th>
<th>14.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
<td></td>
</tr>
</tbody>
</table>

BLOSUM 62 scores for residue pairs

<table>
<thead>
<tr>
<th>A</th>
<th>0.0</th>
<th>4.24</th>
<th>8.34</th>
<th>14.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.0</td>
<td>8.84</td>
<td>10.78</td>
<td></td>
</tr>
</tbody>
</table>

6. Sorted seed list

| 1 | 2 | 3 | 4 | 5 |

7. Generate seed alignments

Entries 1, 2 and 4 yield a seed alignment.
• Scores between residue pairs
• Descending order sorted pair-scores
• Selection of top pair from left moving right on the string

Algorithm 1 Generation of seed alignments

\[
\text{for } i = 1 \text{ to } m \times n \text{ do} \\
\quad \text{for } j = i \text{ to } m \times n \text{ do} \\
\quad \quad p \leftarrow SMM[j].\text{residue}[0] \\
\quad \quad q \leftarrow SMM[j].\text{residue}[1] \\
\quad \quad \text{if } p \text{ and } q \text{ are not already mapped then} \\
\quad \quad \quad \text{if RMSD criteria is met for } map \cup (p, q) \text{ then} \\
\quad \quad \quad \quad \{ \text{Above check considers current alignment type} \} \\
\quad \quad \quad \quad map \leftarrow map \cup (p, q) \\
\quad \quad \quad \text{end if} \\
\quad \quad \text{end if} \\
\quad \quad \text{end for}\{\#j\} \\
\quad \text{Update database with } map \\
\text{end for}\{\#i\} 
\]
Two binding sites are represented as sets of residues
\[ S = \{R_1 \ldots R_m\} \] where \( R_i \) is \( i^{th} \) residue of first site
Each residue defines a partitioning of the set of atoms, \( A \)
\[ R = \{a \in A\} \subset A \]
\[ R_i \cap R_j = \emptyset (\forall i \neq j \in |S|) \]
Where \(|S|\) denotes cardinality of the set, \( S \)
Similarly second site is represented by \( S' = \{R'_1 \ldots R'_n\} \) on set of atoms, \( A' \)
Chemical similarities are denoted by a function \( BL : S \times S' \rightarrow N \)
Geometric similarities (GPS) are denoted by \( GPS : S \times S' \rightarrow N \)
A combination scoring scheme is defined \( GPS \times BL_{ij} \rightarrow GPS_{ij} \ast BL_{ij} \)
A linearization of \( GPS \times BL \) is performed
A one-to-one function is defined \( L : [1 \ldots m] \times [1 \ldots n] \rightarrow [1 \ldots m \ast n] \)
SeedList is created by obtaining values from \( GPS \times BL \)
\( SeedList^V_{L(i,j)} \leftarrow GPS \times BL_{ij} \) for storing the values
\( SeedList^P_{L(i,j)} \leftarrow (i, j) \) for storing the residue pairs
SeedList is sorted such that \( (\forall p \leq q)SeedList^V_p \geq SeedList^V_q \)
A mapping is defined as residuewise correspondences between the two sites
A one-to-one function, for a mapping \( M : [1 \ldots m] \rightarrow [1 \ldots n] \)
Seed mapping or alignment \( B \) is derived by traversal of SeedList
\( B \leftarrow \{(p, q)\} \subset SeedList^P \)
PocketAlign (Validation & New Results)

• Ran for a set of 34 pairs of sites known to be similar
• Encouraging results obtained from a set of 143 pairs of histamines, 29 pairs of lectins, 209 pairs of sites of carbohydrate (GAL, GLC and MAN) sites and ATP binding sites
Cladogram of MHC class I binding sites of 120 molecules from various species
targetTB – Target Identification Pipeline

(A&B) Systems and Sequence Level Filters
- A1 Node deletions on STRING + Metabolic Influences network
- A2 Essential genes from Mtb iA/J661, GSMN-TB
- A3 High-throughput Transposon Site Hybridisation (TraSH) Mutagenesis study
- B Eliminated proteins with close homologues in human proteome

(C) Structural Assessment of Targetability
- Binding site prediction and comparison – Mtb vs. Hsa
- Structural models obtained from ModBase
- Binding sites identified using PocketDepth and compared using PocketMatch (cut-off: 0.80) (A&B&C ⇒ D)

Other Filters (applied to (D))
- E Expression of target (Microarray data)
- F Non-similarity to human ‘anti-targets’
- G Non-similarity to gut flora proteins
- H Paths to resistance mechanisms

Multiple Lists of Targets
- H Passing filters A–G
- I H-List targets upregulated in persistence
- J H-List targets that can serve as broad-spectrum targets
- K H-List targets, unique to Mtb.

Raman et al.; 2008, BMC Syst. Biol
THE NEW DRUG DISCOVERY PIPELINE

Chandra, Expert Opin Drug Disc. 2009
Thank You