BSAlign: Rapid Detection of Ligand-binding Sites in Protein Structures

Zeyar Aung & Joo Chuan Tong
Data Mining Department
Institute for Infocomm Research, A*Star
Singapore

19th International Conference on Genome Informatics (GIW 2008)

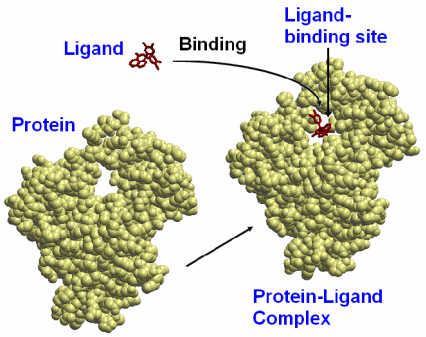
Outline

- Problem Definition
- Motivation
- Existing Methods
- Out Proposed Method: BSAlign
  - Graph Representation
  - Detection of Similar Binding Sites
- Experimental Results
- Future Work and Conclusion

Problem Definition

Proteins and Ligands

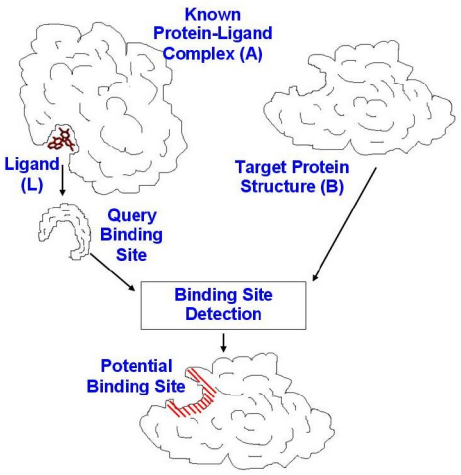
- Protein:
  - Physical basis of life.
  - Performs vital functions in cells.
  - Amino acid sequence folded into 3D structure.



- Ligand:
  - A specific chemical compound that binds to a specific protein to form a complex.
  - Can inhibit, promote, or alter the function of the receptor protein.
- Ligand-binding Site
  - The region in the receptor protein where the ligand binds.

Ligand-binding Site Detection

- Input:
  1. Protein structure *A* with ligand *L* bound to it (i.e. a known protein-ligand complex).
  2. Target protein structure *B* to which ligand *L* is suspected to be bound.
- Output:
  - The potential binding site of *L* in *B* that is similar to that of *L* in *A*.



Motivation

Ligand-binding site detection is useful for:

- Protein function prediction
  - Proteins that binds to the same ligand tends to have similar functions.
- Drug discovery
  - New drug target identification
  - Generation of targeted drug leads like inhibitors
  - Side-effect prediction, etc.

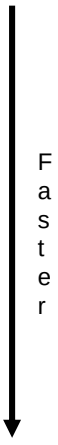
Existing Methods

Generic Structure Comparison Methods

- CE, DALI, SSAP, etc.
  - Purely geometrical; does not take physicochemical properties into account.
  - Sequential alignment.
  - Not suitable for the task of binding site detection.

Dedicated Binding Site Detection Methods

- Graph-based
  - ASSAM (pseudo-atoms, sub-graph isomorphism)
  - eF-site, Cavbase (pseudo-centers, sub-graph isomorphism)
  - SiteEngine (pseudo-centers, geometric hashing)
- Fingerprint-based
  - SiteAlign (cavity fingerprint, fast comparison)



Limitations of existing methods:

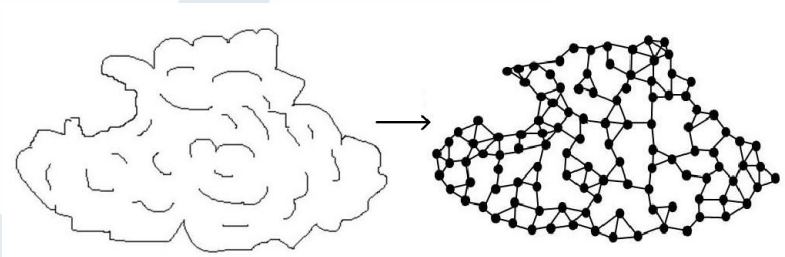
- Graph-based Methods
  - Use sub-residue level of representation (pseudo-atoms or pseudo-centers).
  - Accurate but slow.
- Fingerprint-based Methods
  - Concise representation and time-efficient comparison.
  - Fast, but not accurate.

- Our objective:
  - To develop a ligand-binding site detection method that is both accurate and fast.

Our Proposed Method

BSAlign = Binding Site Aligner

Protein Structure as Graph



- Vertex: each residue is a vertex.
- Edge: two vertices are connected if are at most 15Å apart.

Vertex and Edge Labels

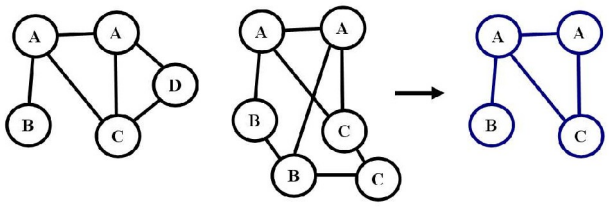
- Vertex labels (for residues)
  1. Solvent accessibility (%)
  2. Physicochemical property (non-polar, polar, aromatic, positive, negative)
  3. Secondary structure type (helix, sheet, loop)

- Edge labels (for relationship between two residues)
  1. Distance between CA atoms of two residues
  2. Angle between CA-CB vectors of two residues

DSSP Algorithm is used to calculate the solvent accessibilities and the secondary structures.

Detection of Similar Binding Sites

- Represent both the query binding site and the target protein structure as graphs.
- Find the largest common portions in two graphs (maximum common sub-graph isomorphism).

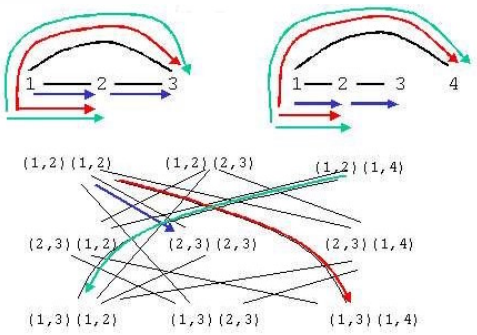


- The larger the common sub-graph the more similar the two input structures are.

Maximum Common Sub-graph Isomorphism

Edge-Product Graph

- Transform the two input graphs (binding site and one for target structure) into a single edge-product graph.
- The resultant edge-product graph depends on the compatibilities of vertices and edges in two input graphs.



Edge-Product Graph

An edge-product graph GP of two input graphs $G1 = (V1, E1)$ and $G2 = (V2, E2)$ is defined as $GP = (VP, EP) = (E1 \times E2)$ in which:

- The vertex set VP of the product graph consists of all the compatible edge pairs in $E1$ and $E2$. That is, $vp_i = (e1_r, e2_s)$ if:
  - $EC(e1_r, e2_s) = \text{TRUE}$, and
  - $(VC(a1_r, a2_s) = \text{TRUE} \wedge VC(b1_r, b2_s) = \text{TRUE}) \vee$ $(VC(a1_r, b2_s) = \text{TRUE} \wedge VC(b1_r, a2_s) = \text{TRUE})$
- There exists an edge between two vertices $vp_i = (e1_r, e2_s)$ and $vp_j = (e1_t, e2_u)$ of the product graph if:
  - $(e1_r \neq e1_t) \wedge (e2_s \neq e2_u)$, and
  - Either:
    - $(e1_r$ and $e1_t$ have a common vertex $v1_{rt}) \wedge (e2_s$ and $e2_u$ have a common vertex $v2_{su}) \wedge (VC(v1_{rt}, v2_{su}) = \text{TRUE})$, or
    - ($e1_r$ and $e1_t$ do not have a common vertex) $\wedge$ ($e2_s$ and $e2_u$ do not have a common vertex)

Vertex and Edge Compatibilities

$$VC(v_i, v_j) = \begin{cases} \text{TRUE} & \text{if } (|v_i.SA - v_j.SA| \leq T1_{SA}) \vee \\ & \quad ((|v_i.SA - v_j.SA| \leq T2_{SA}) \wedge \\ & \quad (v_i.PT = v_j.PT) \wedge (v_i.SS = v_j.SS)) \\ \text{FALSE} & \text{otherwise} \end{cases}$$

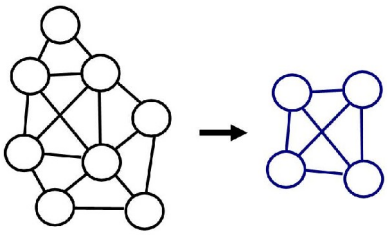
SA = Solvent accessibility; PT = Physicochemical type;
SS = Secondary structure type

$$EC(e_i, e_j) = \begin{cases} \text{TRUE} & \text{if } ((|e_i.DC - e_j.DC| \leq T_{DC}) \wedge \\ & \quad (|e_i.AN - e_j.AN| \leq T_{AN})) \\ \text{FALSE} & \text{otherwise} \end{cases}$$

DC = Distance between CA atoms;
AN = Angle between two CA-CB vectors

Maximum Clique Detection

- Find the maximum clique (fully-connected sub-graph) in the edge-product graph.
- The maximum clique corresponds to the maximum common sub-graph.
- CLIQUER algorithm (Helsinki University of Technology) is used.



Auto-tuning of Parameters

- Clique detection is a time-consuming process.
- If the edge-product graph is too big, the detection of the maximum clique in it is very slow.
- Iteratively re-construct the edge-product graph with stricter vertex and edge compatibility parameters each time.
- Stop when the number of edges is in the edge-product graph becomes less than 1 million.

Mapping of Matching Edge Pairs into Matching Vertex Pairs

- Each vertex in the maximum clique of the edge-product graph is a pair of edges of the original graphs that are matching.
- Matching edge pairs are mapped into matching vertex pairs (i.e. alignment of residues) using the Hungarian algorithm for optimal assignment.

Matching Edge Pairs (query)		(target)
1, 2	–	53, 55
1, 8	–	51, 53
2, 3	–	55, 57
3, 4	–	57, 60
4, 5	–	58, 60
7, 9	–	54, 56

⇒

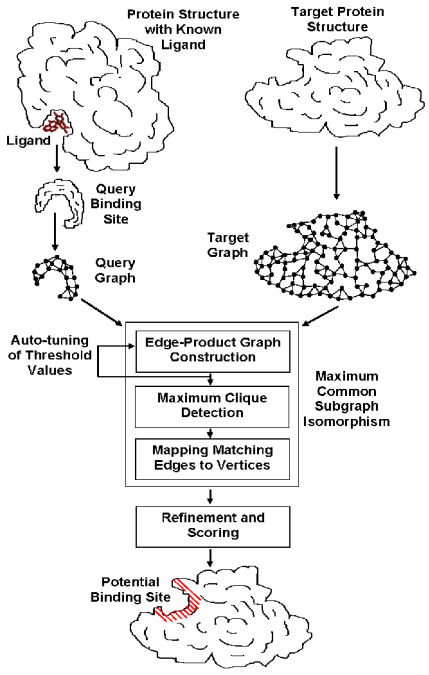
Matching Vertex Pairs (query)		(target)
1	–	53
2	–	55
8	–	51
3	–	57
4	–	60
5	–	58
7	–	54
9	–	56

Refinement and Scoring

- Scoring Function:

$$\text{Alignment score} = \frac{3 \times \text{No. of aligned residues}}{1 + \text{RMSD}}$$

- RMSD (root-mean square deviation) is calculated by superimposing the matching (aligned) residues of one structure onto another.
- The initial alignment is iteratively refined by removing the furthest residue pair form the superimposed structures and then re-superimposing the remaining structures.
- Stop the refinement when the alignment score cannot be further improved.



Experimental Results

Data Set

- Shulman-Peleg *et al.* (J. Mol. Biol., 2004)
- 126 proteins: 34 adenine-binding proteins + 92 other proteins

Functional Type	Total	SCOP Folds	PDB IDs
Adenine-binding proteins	34	18	1a49, 1a82, 1ads, 1atp, 1ayl, 1b4v, 1b8a, 1bx4, 1byq, 1csc, 1csn, 1e2q, 1e8x, 1f9a, 1fmw, 1g5t, 1gn8, 1hck, 1hpl, 1j7k, 1jjv, 1kay, 1kp2, 1kpf, 1mjh, 1mmg, 1nhk, 1nsf, 1phk, 1qmm, 1yag, 1zin, 2src, 9ldt
Other proteins	92	21	1a27, 1a52, 1abi, 1acb, 1alq, 1arb, 1azm, 1b56, 1b6o, 1bt5, 1cbs, 1cho, 1com, 1cqq, 1cse, 1csm, 1dbf, 1dcs, 1e6w, 1ecm, 1ela, 1elc, 1equ, 1ere, 1err, 1exm, 1fby, 1fds, 1fem, 1flj, 1fnj, 1fnk, 1ftp, 1g5y, 1ghp, 1gx9, 1hah, 1har, 1hms, 1hne, 1hsg, 1hsh, 1hwr, 1ifc, 1jd0, 1jgl, 1keq, 1kop, 1kqw, 1kzk, 1l2i, 1lhu, 1lib, 1lid, 1lie, 1lvo, 1mbm, 1mdc, 1mml, 1mu2, 1oh0, 1opa, 1opb, 1pek, 1pmp, 1ppf, 1pro, 1q2w, 1qjg, 1qkt, 1rxf, 1sbn, 1sga, 1sgc, 1tgs, 1tyr, 1vrt, 1whs, 1ysc, 1znc, 2alp, 2cbr, 2ifb, 2lbd, 2lpr, 3ert, 3prk, 3sga, 3tec, 4csm, 4sgb, 4tgl
Total	126		

Experiment

- Query binding site: ATP-binding sites of protein 1atp.
- Target proteins: 126 proteins.
- Compare the query binding site against the target proteins one-by-one.
- Rank those 126 proteins by their alignment scores.

15 Highest Ranking Proteins

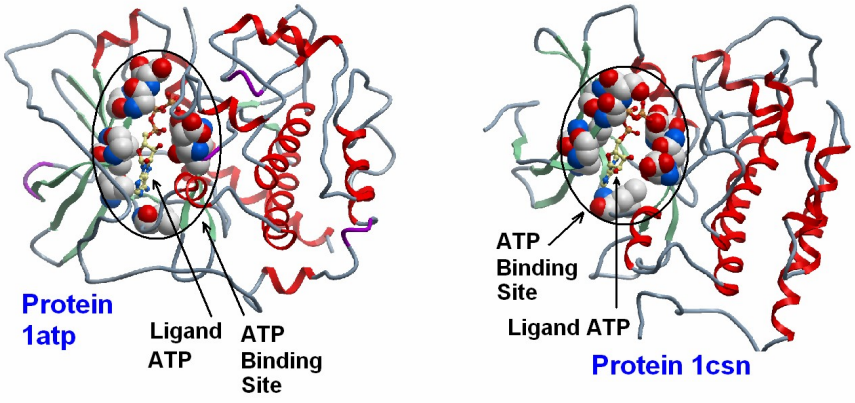
Rank	PDB ID	Protein Name	SCOP Fold Name	Sequence Identity (%) [a]	Aligned Residues	RMSD (Å)	Align -ment Score	Ligand	Functional Type
1	1atp	cAMP-dependent PK, catalytic subunit	Protein kinase-like (PK-like)	100.0	13	0.00	39.00	ATP	Adenine-binding
2	1csn	Casein kinase-1, CK1, catalytic subunit	Protein kinase-like (PK-like)	17.0	10	0.48	20.24	ATP	Adenine-binding
3	2src	c-src protein tyrosine kinase	SH3-like barrel	13.4	11	0.97	16.74	ANP	Adenine-binding
4	1phk	gamma-subunit of glycogen phosphorylase kinase (Phk)	Protein kinase-like (PK-like)	24.2	8	0.58	15.18	ATP	Adenine-binding
5	1hck	Cyclin-dependent PK, CDK2	Protein kinase-like (PK-like)	19.5	9	1.06	13.12	ATP	Adenine-binding
6	3prk	Proteinase K	Subtilisin-like	2.5	6	1.10	8.55	MSU	other
7	1jd0	Carbonic anhydrase	Carbonic anhydrase	4.2	6	1.44	7.37	AZM	other
8	1mjh	"Hypothetical" protein MJ0577	Adenine nucleotide alpha hydrolase-like	15.4	6	1.47	7.27	ATP	Adenine-binding
9	1fnk	Chorismate mutase	Bacillus chorismate mutase-like	7.3	6	1.48	7.25	CSD	other
10	1zin	Adenylate kinase	P-loop containing nucleoside triphosphate hydrolases	10.1	6	1.53	7.13	AP5	Adenine-binding
11	1abi	Thrombin	Trypsin-like serine proteases	16.4	6	1.75	6.53	HMR	other
12	1hah	Eukaryotic proteases	Trypsin-like serine proteases	16.7	6	1.76	6.52	NAG	other
13	1kp2	Argininosuccinate synthetase	Adenine nucleotide alpha hydrolase-like	8.8	6	1.78	6.47	ATP	Adenine-binding
14	1dbf	Chorismate mutase	Bacillus chorismate mutase-like	3.7	6	1.79	6.45	SO4	other
15	1nsf	Hexamerization domain of N-ethylmalemide-sensitive fusion (NSF) protein	P-loop containing nucleotide triphosphate hydrolases	12.4	6	1.81	6.41	ATP	Adenine-binding

Interpretation of Results

- 9 out of 15 (60%) of the highest ranking proteins are the adenine-binding proteins (bound to adenine-containing compounds like ATP, ANP, and AP5).
- BSAlign provides same level of accuracy as a state-of-the-art method called SiteEngine (Shulman-Peleg *et al.*, J. Mol. Biol., 2004).
- Among these 9 adenine-binding proteins, 8 are also reported by SiteEngine.

Example

- ATP-binding sites of 1atp and 1csn:
  - No. of aligned residue = 10; RMSD = 0.48Å.



Running Time

- Platform: PC with Pentium D 3.2GHz CPU and 2GB running Linux.
- 1atp x 126 proteins
  - BSAlign: 871 seconds (14 minutes and 31 seconds)
  - SiteEngine: 12,010 seconds (3 hours, 18 minutes, and 10 seconds)
- BSAlign is about 14 times faster than SiteEngine while offering the same level of accuracy.

Conclusion and Future Works

Conclusion

- Ligand-binding site detection is an important problem in protein function prediction and drug discovery.
- We have proposed a new binding-site detection method called “BSAlign”.
- BSAlign employs a residue-based graph-representation and maximum common sub-graph isomorphism to detect the similarity of the binding sites.
- BSAlign is 14 times faster than a state-of-the-art method called SiteEngine while offering the same level of accuracy.
- The efficiency contribution of BSAlign can be very useful for speed-critical applications like drug discovery.

Future Works

- To evaluate BSAlign with a larger and more diverse data set in order to ascertain its accuracy performance.
- To try out a different sub-graph isomorphism model by using the vertex-product graphs.

Acknowledgement

- Dr Joo Chuan Tong (PI, Molecular Design Group, Institute for Infocomm Research)
- Dr See-Kiong Ng (Department Head, Data Mining Department, Institute for Infocomm Research)

- Thank you all for listening.
- Any questions?