Appendix B: Similarity Score Calculations

Below we present the details of the implementation of the SiteEngine algorithm and of its scoring functions.

General Data Structures and Notations Let M and B denote the sets of atomic coordinates of a whole protein structure and of a query binding site. Let M_S and B_S denote their surfaces (defined by sets of 3-D surface points) respectively. The default threshold for the maximal distance is denoted by ϵ =3.0Å.

- *Distance Transform Grid* - is a 3D grid, in which each voxel holds a value corresponding to the distance to the surface of the molecule. There are three types of voxels, corresponding to the interior, exterior and surface of the protein. The definition and implementation of a distance transform grid is according to Duhovny et al.(0). Physico-chemical labeling has been added to allow efficient matching of the properties.
- $DT\_dist(M, p)$ - denotes the distance of a 3-D point p to the surface of a molecule M stored in the Distance Transform (DT) Grid.
- $chem(p)$ - denotes the physico-chemical property of the point p. The properties that can be assigned are: Hydrogen bond donor, Hydrogen bond acceptor, Hydrogen bond donor/acceptor, Aliphatic Hydrophobic, Aromatic (pi contacts). Assignment of the properties to surface points, are according to the properties of the corresponding atoms. Surface points created by several atoms with different properties are left unassigned.
- $DT\_chem(M, p)$ - denotes the physico-chemical labeling of the grid voxel to which p belongs. The voxel is marked according to the property of an atom with the largest radius stored in that voxel. The properties are the same as for $chem(p)$.
- $charge(p)$ - denotes the charge of the point p. The charge is assigned according to the side chain to which p belongs. Side-chains of Arg, Lys and His are considered to be positively charged, whereas those of Asp and Glu are negatively charged.
- $DT\_charge(M, p)$ - denotes the charge of the grid voxel to which p belongs.
- $shape(p)$ - denotes the solid angle shape function(0) calculated at point p.

- $DT\_shape(M, p)$ - denotes the shape function of the surface patch, which is created by the physico-chemical property stored in the same grid voxel as p.
- *density* - denotes the density of the Connolly surface representation, which is defined by the number of surface points in $1\mathring{A}^2$ (the default is 10).
- T - denotes a 3D transformation (rotation and translation) that superimposes the query binding site upon the database molecule.

Low-Resolution Scoring Let $dist(p) = |DT\_dist(M, T(p))|$. Let P denote a set of *patch centers* of the query binding site for which $dist(p) \leq 3.0\mathring{A}$. Let $P_{ALI} \subseteq P$, $P_{PI} \subseteq P$ and $P_{HB} \subseteq P$ denote the points of P with aliphatic hydrophobic, aromatic and H-bonding properties respectively for which $chem(p) \simeq DT\_chem(M, T(p))$.
The low resolution score will be calculated in the following way:

$$chem\_score(p) = \begin{cases} 0, \ chem(p) \neq DT\_chem(M, T(p)) \\ 1, \ p \in P_{PI} \\ 1, \ p \in P_{HB} \ \wedge \ charge(p) \neq DT\_charge(M, T(p)) \\ 2, \ p \in P_{HB} \ \wedge \ charge(p) = DT\_charge(M, T(p)) \\ 1/(1 + |DT\_shape(M, T(p)) - shape(p)|), \ p \in P_{ALI} \end{cases}$$

$$Low\_Resolution\_Score(T) = \sum_{p \in P} (1 + chem\_score(p)) \cdot (\epsilon - dist(p)) \quad (1)$$

Overall Surface Score Calculations We apply the transformation T to the query binding site (B) and partition its surface points according to their distance to the surface of the database molecule (M). We distinguish between three distance layers S_0, S_1, S_2 defined in the following way:
$\forall 0 \leq i \leq 2 \ S_i = \{p \in B_S || DT\_dist(M, T(p))| \leq i\}$
At these layers we identify points that in addition to the distance requirements possess similar physico-chemical properties and charges. The charge is compared only for points with the same H-bonding property. We denote these point sets as P_0, P_1, P_2 respectively:
$\forall 0 \leq i \leq 2 \ P_i = \{p \in S_i | chem(p) \simeq DT\_chem(M, T(p))\}$
In addition we consider the charges of the exposed to the surface H-bonding properties:
$C_0 = \{p \in P_0 | charge(p) = DT\_charge(M, T(p))\}$
Then the overall surface score is defined as:

$$Overall\_Surface\_Score(T) = 1/density \cdot \sum_{i=1}^{i=2}(\epsilon - i)(|S_i| + |P_i|) + |C_0| \quad (2)$$

Match List (1:1 Correspondence) Definition The match list is defined by calculating the maximum weight matching in a bipartite graph(0; 0). Given two sets of pseudocenters P and Q of the molecule and of the binding site and given a transformation T (rotation and translation), the task is to find the largest set of points pairs,$\{(p_1, q_1)...(p_n, q_n)\}$, so that the points of each pair are most similar in their geometrical and physico-chemical properties. The maximal allowed distance between a pair of matched pseudocenters is ϵ =3.0Å. We solve the matching problem by means of a bipartite graph $G(V, E)$ constructed in the following way:

- The nodes of the graph (V) are the pseudocenters of the two molecules. $V = P \cup Q$.
- An edge ($e \in E$) is added between each pair of pseudocenters p_i and q_i for which
  $\parallel p_i - q_i \parallel \leq dist\_thr \ \wedge \ |shape(p_i) - shape(q_i)| \leq shape\_thr \ \wedge \ chem(p_i) \simeq chem(q_i)$.
  Let $E_{ALI} \subseteq E$, $E_{PI} \subseteq E$ and $E_{HB} \subseteq E$ denote the edges connecting the nodes with aliphatic hydrophobic, aromatic and H-bonding properties respectively. Each edge is assigned a weight in the following manner:
  $$weight(e) = \begin{cases} 1/(1.0+ \parallel p_i - T(q_i) \parallel), \ e \in E_{PI} \\ 1/(1.0+ \parallel p_i - T(q_i) \parallel), \ e \in E_{HB} \ \wedge \ charge(p_i) = charge(q_i) \\ 1/(1.5+ \parallel p_i - T(q_i) \parallel), \ e \in E_{HB} \ \wedge \ charge(p_i) \neq charge(q_i) \\ 1/(1.0+ \parallel p_i - T(q_i) \parallel + 2 * (shape(p_i) - shape(q_i))), \ e \in E_{ALI} \end{cases}$$

A match in graph G is a subset pf edges $\hat{E} \subseteq E$ so that no two of them share an endpoint. A node $v \in V$ is called matched with respect to $\hat{E}$ if there is an edge in $\hat{E}$ incident to v. The maximum weight matching of a bipartite graph(0) will therefore represent the largest set of point pairs, which are most similar in their physico-chemical and geometrical properties.

Scoring of Matched Patches Let P and Q denote the sets of pseudocenters of the molecule M and of a binding site B respectively. At the previous stage of match list definition we have obtained a transformation T and a correspondence $\{(p_1, q_1)...(p_n, q_n)\}$ between subsets of the pseudocenters P

and Q. Let $w(p_i, q_i) = \begin{cases} 1, & charge(p_i) = charge(q_i) \\ 0, & otherwise \end{cases}$
First we calculate a score of the spatial similarities between the matched centers.

$$Distance\_Score(T) = \sum_{i=1}^{n}(1 + w(p_i, q_i)) \cdot (\epsilon - \parallel p_i - T(q_i) \parallel) \quad (3)$$

Note: In the output files of the SiteEngine package this score is entitled: "1:1 correspondence distance score".

In addition we estimate the similarity between the corresponding surface patches of Aliphatic Hydrophobic and Aromatic properties. Let $S_{p_i} = \{p_i^s\}$ and $S_{q_i} = \{q_i^s\}$ denote the surface patches created by atoms contributing to the properties of pseudocenters p_i and q_i. The mutual overlap between these patches is calculated as defined by Schmitt et al.(0):
$R_{p_i}^{q_i} = \{p_i^s \in S_{p_i} | \parallel p_i^s - T(q_i^s) \parallel \leq 1.0\mathring{A}\}$
$R_{q_i}^{p_i} = \{q_i^s \in S_{q_i} | \parallel q_i^s - T^{-1}(p_i^s) \parallel \leq 1.0\mathring{A}\}$
We define the size of the mutual overlap by:
$Overlap\_Size(S_{p_i}, S_{q_i}) = min(|R_{p_i}^{q_i}|, |R_{q_i}^{p_i}|)$
We estimate the shape of the overlap by calculating the Connolly shape function(0) in a sphere bounding the smallest overlap $R_m = min(R_{p_i}^{q_i}, R_{q_i}^{p_i})$. Let V_{p_i} and V_{q_i} denote the shapes of the patches of p_i and q_i respectively. The score of the overlap is calculated in the following manner:

$$Curvature\_Score(T) = \sum_{i=1}^{n}\{1 + (Overlap\_Size(S_{p_i}, S_{q_i})/[density \cdot (1 + 10 * |V_{p_i} - V_{q_i}|)]\} \quad (4)$$

Note: In the output files of the SiteEngine package this score is entitled: "1:1 correspondence curvature score".

The Total Score The final (total) score is the sum of all the scores calculated by the program:

$$Total\_Score(T) = Overall\_Surface\_Score(T) + \\ + density \cdot [Low\_Resolution\_Score(T) + Match\_Score(T) + Curvature\_Score(T)] \quad (5)$$

Note: At the SiteEngine web server site this score is entitled "Similarity Score".

References

Duhovny, D., Nussinov, R. & Wolfson, H. (2002). Efficient unbound docking of rigid molecules. In *Workshop on Algorithms in Bioinformatics*, (Guigo, R. & Gusfield, D., eds), vol. 2452, pp. 185–200. Springer Verlag.

Connolly, M. L. (1986). Measurement of protein surfaces shape by solid angles. *J. Mol. Graph.* **4**, 3–6.

Mehlhorn, K. (1999). *The LEDA platform of combinatorial and geometric computing*. Cambridge University Press.

Cormen, T. H., Leiserson, C. E. & Rivest, R. L. (1990). *Introduction to Algorithms*. The MIT Press.

Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence or fold homology. *J. Mol. Biol.* **323**, 387–406.

Connolly, M. (1986). Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers,* **25**, 1229–1247.