Appendix B: Similarity Score Calculations

We define an *interface* as an unordered pair of interacting binding sites (A and B), that belong to two non-covalently linked protein molecules. Two interfaces are considered to be similar, if the binding sites that comprise them share similar physico-chemical properties and shapes. Given two interfaces I=(A, B) and I'=(A', B') the goal is to find the best alignment between them. Specifically, let S denote a set of scoring functions that are used to measure the similarity of the aligned properties. The problem which is heuristically solved by I2I-SiteEngine can be formalized as follows: find a rigid transformation T (rotation and translation) that maximizes the value of $S(I, T(I'))$. Below we provide the details of the definition of the scoring functions S.

In addition, we assume that the correspondence between the binding sites of the two complexes is unknown, meaning that the binding site A, can be aligned either to A' or to B'. Since the algorithmic procedures that are applied in both cases are the same, the description below refers only to the first option. However, both alignments are considered by the method and the solution that provides the highest score is selected.

General Data Structures and Notations Let ϵ denote the threshold for the maximal distance between matched pseudocenters from the corresponding binding site of the compared interfaces. The default value is 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.
- $DT\_dist(A, p)$ - denotes the distance of a 3-D point $p \in A'$ 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.

- $DT\_chem(A, p)$ - denotes the physico-chemical labeling of the grid voxel to which $p \in A'$ 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.
- $DT\_charge(A, p)$ - denotes the charge of the grid voxel to which $p \in A'$ belongs.
- $shape(p)$ - denotes the solid angle shape function(0) calculated at point p.
- $DT\_shape(A, p)$ - denotes the shape function of the surface patch, which is created by the physico-chemical property stored in the same grid voxel as $p \in A'$.
- *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 $I=(A, B)$ and $I'=(A', B')$ be the compared interfaces.For $p \in A'$ let $dist(p, M) = |DT\_dist(M, T(p))|$ denote the distance of the point p of A' from the surface of A after the superimposition. Let P (or $\hat{A}$ and $\hat{B}$) denote a set of *patch centers* of the binding sites A' or B' for which $dist(p, M) \leq 3.0\mathring{A}$, where M is the binding sites of A and B respectively. 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 $chem(p, M) \simeq DT\_chem(M, T(p))$.
The low resolution score will be calculated in the following way:

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

$$Low\_Resolution\_Score(T) = \sum_{p \in \hat{A}} (1 + chem\_score(p, A)) \cdot (\epsilon - dist(p, A)) + \sum_{q \in \hat{B}} (1 + chem\_score(q, B)) \cdot (\epsilon - dist(q, B)) \quad (1)$$

Overall Surface Score Calculations We apply the transformation T to the query interface $I{=}(A, B)$ and partition its surface points according to their distance to the surface of the database interface $I'{=}(A', B')$. We distinguish between three distance layers S_0, S_1, S_2 defined in the following way:
$\forall 0 \leq i \leq 2\ S_i = \{p \in A' || DT\_dist(A, T(p))| \leq i + \{p \in B' || DT\_dist(B, T(p))| \leq i\}$
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, and let M be the binding sites of either A' or B':
$\forall 0 \leq i \leq 2\ P_i = \{p \in S_i | chem(p) \simeq DT\_chem(M, T(p))\}$
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). A set of pseudocenters of an interface $I{=}(A, B)$ is the union of the sets of pseudocenters of binding sites A and B that constitute it. The correspondence is obtained by calculating the maximum weight match in a weighted *bipartite graph*(0; 0), which represents the largest set of pairs of similar pseudocenters. The construction of a weighted *bipartite graph* for comparison between two protein-protein interfaces $I{=}(A, B)$ and $I'{=}(A', B')$ is performed in the following way: (1) Each pseudocenter from binding sites A, B, A' and B' defines a node. (2) Assuming that a candidate transformation aligns a binding site A to A' and a binding site B to B', edges of a *bipartite graph* can only connect nodes of A to A' and nodes of B to B'. Following these restrictions, an edge ($e \in E$) is added between each pair of pseudocenters p_i

and q_i for which:
$\| p_i - q_i \| \leq dist\_thr \ \wedge \ |shape(p_i) - shape(q_i)| \leq shape\_thr \ \wedge \ chem(p_i) \simeq chem(q_i)$.
(3) Each edge is assigned a weight that reflects the differences in distance and shape between the nodes. Specifically, 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:

$$weight(e) = \begin{cases} 1/(1.0 + \| p_i - T(q_i) \|), \ e \in E_{PI} \\ 1/(1.0 + \| p_i - T(q_i) \|), \ e \in E_{HB} \ \wedge \ charge(p_i) = charge(q_i) \\ 1/(1.5 + \| p_i - T(q_i) \|), \ e \in E_{HB} \ \wedge \ charge(p_i) \neq charge(q_i) \\ 1/(1.0 + \| p_i - T(q_i) \| + 2 * (shape(p_i) - shape(q_i))), \ e \in E_{ALI} \end{cases}$$

The maximum weight match(0) in this graph, provides a 1:1 correspondence between subsets of pseudocenters of the two interfaces. Due to restriction on the creation of the edges we obtain two separate 1:1 correspondences: one between subsets of pseudocenters of A and A', and another between subsets of B and B'.

Scoring of Matched Patches Let P and Q denote the sets of pseudocenters of the interfaces I and I' 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.

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

Note: In the output files of the I2I-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):
$R_{p_i}^{q_i} = \{p_i^s \in S_{p_i} | \ \| \ p_i^s - T(q_i^s) \ \| \leq 1.0\mathring{A}\}$
$R_{q_i}^{p_i} = \{q_i^s \in S_{q_i} | \ \| \ q_i^s - T^{-1}(p_i^s) \ \| \leq 1.0\mathring{A}\}$
We define the size of the mutual overlap by:

$Overlap\_Size(S_{p_i}, S_{q_i}) = min(|R^{q_i}_{p_i}|, |R^{p_i}_{q_i}|)$
We estimate the shape of the overlap by calculating the Connolly shape function(0) in a sphere bounding the smallest overlap $R_m = min(R^{q_i}_{p_i}, R^{p_i}_{q_i})$. 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 I2I-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 I2I-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.