Weighted geometric docking: Incorporating external information in the rotation-translation scan

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10391 · Source: PubMed

2 AUTHORS:

Efrat Ben-Zeev
Weizmann Institute of Science
14 PUBLICATIONS **270** CITATIONS
SEE PROFILE

Miriam Eisenstein
Weizmann Institute of Science
164 PUBLICATIONS **5,872** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 52:24–27 (2003)

Weighted Geometric Docking: Incorporating External Information in the Rotation-Translation Scan

Efrat Ben-Zeev[1] and Miriam Eisenstein[2]*
[1]*Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot, Israel*
[2]*Department of Chemical Services, The Weizmann Institute of Science, Rehovot, Israel*

ABSTRACT **Weighted geometric docking is a prediction algorithm that matches weighted molecular surfaces. Each molecule is represented by a grid of complex numbers, storing information about the shape of the molecule in the real part and weight information in the imaginary part. The weights are based on experimental biochemical and biophysical data or on theoretical analyses of amino acid conservation or correlation patterns in multiple-sequence alignments of homologous proteins. Only a few surface residues on either one or both molecules are weighted. In contrast to methods that use postscan filtering based on biochemical information, our method incorporates the external data in the rotation-translation search, producing a different set of docking solutions biased toward solutions in which the up-weighted residues are at the interface. Similarly, interactions involving specified residues can be impeded. The weighted geometric algorithm was applied to five systems for which regular geometric docking of the unbound molecules gave poor results. We obtained much better ranking of the nearly correct prediction and higher statistical significance when weighted geometric docking was used. The method was successful even when the weighted portion of the surface corresponded only partially and approximately to the binding site. Proteins 2003;52:24–27.** © 2003 Wiley-Liss, Inc.

Key words: molecular recognition; weighted surface; sequence analysis; conserved patches; mutations; grid representation by complex numbers

INTRODUCTION

Many protein–protein docking methods have been developed in the past few years and used to predict interactions between biological molecules. Most of these methods treat the proteins as rigid bodies and use shape complementarity as well as electrostatics, desolvation, and hydrogen bonding to identify the nearly correct structure and to distinguish it from false solutions. However, biochemical information, such as the residues implicated in the activity of the protein, can and should be used in the prediction procedure. For example, in the prediction of the complex between β-lactamase and BLIP the groups of Eisenstein and Katchalski-Katzir, Janin, Cherfils and Zimmerman, Shoichet and Kuntz and Duncan, Rao and Olson used biological information (knowledge of the active site residues of β-lactamase) to filter the docking solutions.[1] Such filtering was later made an integral part of the prediction procedure.[2] It proved to be very beneficial, eliminating many of the false-positive solutions produced in the full rotation-translation scan and elevating the rank of the nearly correct solution. However, in some cases, the filtering did not help, possibly because the nearly correct solution was not in the list of solutions produced by the scan. This situation is not unrealistic in methods that use explicit rotations but scan the translations via fast Fourier transformations and save only a few high scoring solutions per orientation (often only one solution is saved). Hence, if false-positive solutions are formed at the orientation appropriate for the correct solution, the latter is overlooked, and it does not appear in the list of solutions produced in the scan.

In this manuscript we present a method that incorporates external information in the rotation-translation scan, giving more weight to intermolecular contacts that involve specified residues in either one of the molecules or both. In this way, we generate a different set of solutions, biased toward solutions in which the specified residues are involved (or not involved if this is the preferred option) in binding. We exploit the vector nature of complex numbers, storing information about the shape of the molecule in the real part and weight information in the imaginary part. Previously, we used a similar approach to include electrostatic complementarity in docking, storing information about the electrostatic character of the molecule in the imaginary part[3]; currently, we are testing the importance of hydrophobic complementarity in the same manner (Alexander Berchanski and Miriam Eisenstein, manuscript in preparation). In the future, shape complementarity, electrostatic complementarity, hydrophobic complementarity, and external information will be combined to provide a more comprehensive description of the docked molecules.

The weighted geometric docking method was tested on five enzyme-inhibitor and antibody-antigen systems, improving significantly the docking results in every case. It

*Correspondence to: Miriam Eisenstein, Department of Chemical Services, The Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: miriam.eisenstein@weizmann.ac.il

Received 28 October 2002; Accepted 16 December 2002

TABLE I. Comparison of the Results of Weighted Geometric and Geometric Docking of Unbound Structures

System	Geometric docking			Weighted geometric docking			Interface SAS[a]; up-weighted SAS (Å^2)	References
	Rank[b]	Score	ΔZ[c]	Rank	Score	ΔZ		
Acetylcholinesterase/Fasciculin-II 1fss:2ace/1fsc	689–705	435	−5.7	21–22	630	−2.1	907.6; 468.1	(9–11)
Subtilisin/eglin-C 2sec:1scd/1tec	918–949	423	−7.0	1	936	0.0	650.6; 578.5	(12–14)
Ab D1.3/Ab E5.2 1dvf:1vfa/1dvf[d]	2725–2793	417	−5.7	128–134	570	−3.0	669.4; 241.3	(15,16)
Ab D1.3/lysozyme 1vfb:1vfa/1lza	Not found			492–506	464	−6.7	611.9; 272.3	(16,17)
Ab Jel42/HPr 2jel[e]:2jel[d]/1poh	5270–5424	346	−7.1	359–371	467	−4.3	661.8; 177.3	(18,19)

[a]This column lists the solvent-accessible surface area (SAS) of molecule **a,** which is buried on complex formation, and the SAS of the up-weighted residues. These values were calculated from the experimental structures of the complexes.
[b]The ranks are given as a range of numbers when the score of the nearly correct solution is equal to the scores of other solutions.
[c]ΔZ is the difference between the Z values of the nearly correct solution and the top ranking solution (in σ units). We find that ΔZ represents the improvement in the docking results better than Z, because it compares the nearly correct solution to the best false-positive solution, whereas Z compares it to the average false solution.
[d]The unbound structure of this molecule was not available; therefore, the bound structure was used.
[e]The rank and score for this system differ from previously published values[3] because here only the Fv fragment of the antibody was used in the docking scan.

was then applied in the CAPRI experiment[4] and in the docking of colicin E3 to the 30S ribosomal subunit.[5]

MATERIALS AND METHODS

Weighted Geometric Docking

The molecules to be docked are projected onto a three-dimensional (3D) grid of complex numbers. The shape of the molecule is depicted in the real part as previously described.[6] Thus, for molecule **a**, surface grid points are given the value 1, those in the interior are given the value −15, and grid points outside the molecule are given the value 0. For molecule **b**, grid points on the surface and in the interior of the molecule are given the value 1.

In regular geometric docking, the imaginary part of all the grid points is 0. In the weighted geometric docking, it remains 0 for most of the grid points on the surface of molecule **a**. Non-zero values, reflecting the weight, are given only to the imaginary part of grid points derived from the side-chain atoms of a limited selection of specified residues. The weighting scheme depends on whether weight is given to residues of molecule **a**, molecule **b**, or both. To up-weight, or down-weight only specified residues of molecule **a**, we give a non-zero value t to the imaginary part of grid points derived from these residues, as follows:

$$\bar{a}_{l,m,n} = \begin{cases} 1 + it & \textit{for surface grid points derived} \\ & \textit{from the specified residues} \\ 1 & \textit{for other surface grid points.} \end{cases} \quad (1)$$

The grid representation of molecule **b** is:

$$\bar{b}_{l,m,n} = 1 - i \quad \textit{for surface and interior grid points.} \quad (2)$$

A similar modification allows weighing only residues on the surface of molecule **b**. When the two grid representations are correlated, the overlap of a weighted grid point with any surface grid point of the other molecule contributes $1 + t$ to the real part of the correlation value. The contribution of other surface–surface overlaps remains 1. By giving the value $1 - i$ only to grid points derived from specified residues of molecule **b** (and 1 to all the other grid points), only contacts that involve specified residues on both molecules are weighted. In every case, the imaginary part of the interior grid points of molecule **a** is kept 0 to prevent interpenetration. The correlation function in the weighted geometric docking is computed by using fast Fourier transformations, as described before.[6]

The parameter t can be either positive to up-weight certain contacts or negative to impede interactions with specified residue. Notably, the relative orientation of the two molecules is not constrained by the weighting, but for each orientation the translations that form more contacts via the specified residues ($t > 0$), or fewer contacts ($t < 0$), are preferred.

Rotation-Translation Scans

Rotation-translation scans were performed for each of the five systems listed in Table I. The grid interval ranged from 1.1 to 1.2 Å, and the angular interval was 12°, resulting in 8760 relative orientations, and we kept one highest scoring structure per orientation. All the solutions were sorted by their scores, and the nearly correct solution was identified by calculating the root-mean-square deviation (RMSD) between each prediction and the experimental structure (including all the common Cα atoms), requiring a value of ≤3.0 Å.

The observed distribution of scores in each scan was fitted with an extreme value distribution function,[7] providing estimates for the mean score (μ) and the standard deviation (σ) of the distribution, and the statistical significance (Z) of each solution.[3]

RESULTS

We tested the weighted geometric docking algorithm by applying it to five enzyme-inhibitor and antibody-antigen systems, whose structures are known. We chose these particular systems (see Table I) because regular geometric docking of the unbound structures ranked the nearly

correct solution low.[3] We up-weighted only residues on molecule **a**, as follows: In the enzyme-inhibitor systems, we selected 30 residues that line the binding site of subtilisin (PDB[8] code 2sec) and 29 residue in the active site gorge of acetylcholinesterase, including the rim (system 1fss). In the antigen-antibody systems, we up-weighted 10–12 residues in the third complementarity determining region (CDR) in the light and heavy chains (L3 and H3). In the system 1dvf, which consists of two antibody Fv domains, only the L3 and H3 of antibody D1.3 were up-weighted. The value of t was chosen to be large enough so that a significant increase in the score of the nearly correct solution was observed, but not too large to avoid interpenetration of the molecules. The value $t = 0.45$ appeared adequate.

The scores of the nearly correct solutions are significantly higher in the weighted geometric docking in every case (see Table I). This increase is by far larger than the increase in μ, the average score of the false solutions (data not shown). It is accompanied by a considerable improvement in the rank of the nearly correct solution and in its statistical significance. The latter is evident from the ΔZ values for the nearly correct solutions (see Table I), which are shifted by 2.7–7σ toward the high-score ends of the distributions. Clearly, the identification of the nearly correct solution is much more successful when knowledge of the approximate binding site of one of the molecules is incorporated in the scan.

In 2sec, in which 89% of the binding surface of subtilisin was up-weighted, the increase in the score of the nearly correct solution was exceedingly large, and the prediction was more accurate (RMSD of 0.4Å versus 1.0 Å in the geometric scan). In this case, the rank improved from 918–949 to 1. In the other enzyme-inhibitor system, 1fss, only 52% of the interface was up-weighted. The increase in the score was smaller than for 2sec; nevertheless, the rank of the nearly correct solution improved considerably, from 689–705 to 21–22. In the antibody-antigen systems, we up-weighted only residues from the L3 and H3 loops (36%, 45%, and 27% of the binding surface for 1dvf, 1vfb, and 2jel, respectively), and accordingly, the increase in the score was moderate. Again, the rank of the nearly correct solution improved considerably in every case and so did its statistical significance. In particular, the rank for the 1vfa/1lza system improved from "not-found" to 429–506. In this case, a postscan filtering procedure will prove useless because a nearly correct solution is not formed in the geometric docking scan.

In three cases, our selection of up-weighted residues included residues that are not at the interface: 16, 7, and 4 residues of 29, 30, and 11 residues for systems 1fss, 2sec, and 1vfb, respectively. Nevertheless, the rank of the nearly correct solution improved considerably, suggesting that additional false-positive solutions were not formed. It appears that the weighted geometric docking is successful even when the list of up-weighted residues matches the interaction site only partially and approximately.

DISCUSSION

In weighted geometric docking, we match the shapes of two molecules but give more (or less) weight to contacts made by specified portions of the surface. It is possible to up-weight, or down-weight, a part of the surface of molecule **a**, molecule **b**, or both molecules. In addition, it is possible to give different weights to different parts of the surface of one of the molecules, up-weighting some and down-weighting others.

In the Results section, we chose to up-weight portions of the surface of molecule **a** based on our knowledge of the active site of the enzyme or the recognition site of the antibody. This knowledge is often available in real predictions. For example, in our docking experiment of colicin E3 to the 30S ribosomal subunit we used knowledge of the specific cleavage site on the 16S rRNA,[5] and in the CAPRI experiment,[4] we chose to up-weight the CDRs in antibody-antigen targets. In other cases, the effect of point mutations on the binding of the molecules can be used, although in such experiments it is sometimes unclear whether the mutation affects the binding directly or the conformation (or integrity) of the molecule.

The vast amount of sequence data stored in different databases is another source of biological information that can be used to identify residues involved in protein–protein interactions. The connection between sequences and structures is apparent in the work of Lichtarge et al.[20] In their evolutionary trace method, functionally important residues are extracted from sequence conservation patterns in homologous proteins and mapped onto the protein surface, often generating contiguous patches identifying putative functional interfaces. Similarly, the sequence space method[21] and the self-organizing map method[22] analyze multiple-sequence alignments in an attempt to detect residues comprising the regions of binding. We used a method similar to the evolutionary trace method in our prediction procedure for target 1 in CAPRI,[4] identifying conserved patches on the surfaces of the HPr-kinase and HPr molecules and directing the scan by up-weighting interactions between the two patches.

The analysis of correlated mutations provides another source of information about interactions between proteins or between domains of proteins.[23,24] In this context, we also mention the double-mutant cycle method, which can identify residue–residue interactions at the interface.[25] Values of $\Delta\Delta G$, obtained from double-mutant cycles were recently used to guide the docking of interferon to its receptor.[26]

CONCLUSIONS

Weighted geometric docking is a successful procedure that incorporates external data from different sources, such as biochemical and biophysical experiments or theoretical analyzes of sequence data. It is important to emphasize that the method is successful even when our knowledge of the binding site is approximate, when only a part of it is identified and up-weighted, or when a portion of the weighted surface is erroneous, as described in the Results section (and see Ref. 4).

ACKNOWLEDGMENTS

We thank Professor Ephraim Katchalski-Katzir for his support throughout this study.

REFERENCES

1. Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MN. Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nat Struct Biol 1996;3:233–239.
2. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
3. Heifetz A, Katchalski-Katzir E, Eisenstein M. Electrostatics in protein-protein docking. Protein Sci 2002;11:571–587.
4. Ben-Zeev E, Berchanski A, Heifetz A, Shapira B, Eisenstein M. Prediction of the unknown: inspiring experience with the CAPRI experiment. Proteins 2003;52:41–46.
5. Zarivach R, Ben-Zeev E, Wu N, et al. On the interaction of colicin E3 with the ribosome. Biochimie 2002;84:447–454.
6. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
7. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
9. Harel M, Kleywegt GJ, Ravelli RB, Silman I, Sussman JL. Crystal structure of an acetylcholinesterase-fasciculin complex: interaction of a three-fingered toxin from snake venom with its target. Structure 1995;3:1355–1366.
10. Raves ML, Harel M, Pang YP, Silman I, Kozikowski AP, Sussman JL. Structure of acetylcholinesterase complexed with the nootropic alkaloid, (-)-huperzine A. Nat Struct Biol 1997;4:57–63.
11. LeDu MH, Housset D, Marchot P, Bougis PE, Navaza J, Fontecilla Camps C. Structure of fasciculin 2 from green mamba snake venom: evidence for unusual loop flexibility. Acta Crystallogr D 1992;52:87.
12. McPhalen CA, James MN. Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo. Biochemistry 1988;27:6582–6598.
13. Fitzpatrick PA, Ringe D, Klibanov AM. X-ray crystal structure of cross-linked subtilisin Carlsberg in water vs. acetonitrile. Biochem Biophys Res Commun 1994;198:675–681.
14. Gros P, Fujinaga M, Dijkstra BW, Kalk KH, Hol WGJ. Crystallographic refinement by incorporation of molecular dynamics. The thermostable serine protease thermitase complexed with eglin-c. Acta Crystallogr. B 1994;45:488–499.
15. Braden BC, Fields BA, Ysern X, Dall'Acqua W, Goldbaum FA, Poljak RJ, et al. Crystal structure of an Fv-Fv idiotope-anti-idiotope complex at 1.9 A resolution. J Mol Biol 1996;264:137–151.
16. Bhat TN, Bentley GA, Boulot G, Green MI, Tello D, Dall'Acqua W, et al. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. Proc Natl Acad Sci USA 1994;91:1089–1093.
17. Wilson KP, Malcolm BA, Matthews BW. Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. J Biol Chem 1992;267:10842–10849.
18. Prasad L, Waygood EB, Lee JS, Delbaere LT. The 2.5 A resolution structure of the jel42 Fab fragment/HPr complex. J Mol Biol 1998;280:829–845.
19. Jia Z, Quail JW, Waygood EB, Delbaere LT. The 2.0-A resolution structure of Escherichia coli histidine-containing phosphocarrier protein HPr. A redetermination. J Biol Chem 1993;268:22490–22501.
20. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
21. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
22. Andrade MA, Casari G, Sander C, Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol Cybern 1997;76:441–450.
23. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
24. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 2002;48:611–617.
25. Horovitz A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold Des 1996;1:R121–126.
26. Roisman LC, Piehler J, Trosset JY, Scheraga HA, Schreiber G. Structure of the interferon-receptor complex determined by distance constraints from double-mutant cycles and flexible docking. Proc Natl Acad Sci USA 2001;98:13231–13236.