FEBS Letters 584 (2010) 1163–1168

journal homepage: www.FEBSLetters.org

New measures for estimating surface complementarity and packing at protein–protein interfaces

Pralay Mitra, Debnath Pal *

Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India
Supercomputer Education Research Centre, Indian Institute of Science, Bangalore 560 012, India

ARTICLE INFO

Article history:
Received 1 December 2009
Revised 25 January 2010
Accepted 5 February 2010
Available online 12 February 2010

Edited by Gianni Cesareni

Keywords:
Protein–protein interface
Surface complementarity
Interface packing
Method
Measure
Geometric compatibility

ABSTRACT

A number of methods exist that use different approaches to assess geometric properties like the surface complementarity and atom packing at the protein–protein interface. We have developed two new and conceptually different measures using the Delaunay tessellation and interface slice selection to compute the surface complementarity and atom packing at the protein–protein interface in a straightforward manner. Our measures show a strong correlation among themselves and with other existing measures, and can be calculated in a highly time-efficient manner. The measures are discriminative for evaluating biological, as well as non-biological protein–protein contacts, especially from large protein complexes and large-scale structural studies (http://pallab.serc.iisc.ernet.in/nip_nsc).

1. Introduction

Interactions among protein molecules are highly specific and only selected molecules bind to each other in the cell. This selectivity of interaction is called molecular recognition and can be found in ex vivo applications as well, such as in biotechnology and bioprocess engineering. Geometric compatibility of the interaction surfaces is one of the primary requirements for selective binding of molecules. Consequently, a significant number of methods has been developed in a wide range of application areas for useful estimates of geometric compatibility of protein–protein interfaces, such as in structural proteomics, protein engineering, protein design and protein–protein docking. These use approaches such as (i) spherical harmonics [1–3], (ii) atom densities [4,5], (iii) grids [6–8], (iv) tessellation techniques [9–11], (v) void volume [12,13], and (vi) surface density/normals [9,14–16]. Each method has its specific advantages and disadvantages. For example, spherical harmonics based methods use enlarged probe radii (6 Å as in Max and Getzoff, 1988 [2]; or 3.2 Å, Leicester et al., 1988 [1]) to generate a solvent accessible dot surface, or use surface stretching at the reentrant areas, to overcome the limitation of single valued basis functions. These methods are therefore, more apt at shape matching than geometric surface matching. Similarly, atom densities [4] and grid based methods [8] that work in three-dimensional space to assess geometric compatibility of the surface are likely to produce results with implicit error, as surfaces are two-dimensional objects. Many workers have therefore, mix-and-matched useful features from various methods to improve estimation of geometric compatibility at interfaces. For example, the shape complementarity index by Lawrence and Colman (1993) [15] used the surface normals on dots from the molecular surface generated by Connolly [17]. Norel et al. [16] attempted further improvement on this method by selecting critical points on the Connolly surface that represent knobs and holes at the matching surfaces. The method is deemed as highly accurate, but computationally expensive and therefore, has limitations for large-scale applications, such as in protein docking studies.

In this paper, we present two different yet conceptually simple measures for surface complementarity and atom packing, discriminative for evaluating biological and non-biological protein–protein contacts. These measures show a high correlation with other existing methods, and can be computed easily and efficiently from atom information, making it highly suitable for application on large-size complexes, as well as large-scale structural proteomic studies. We propose that they would be useful in estimation of complementarity and packing for any protein–protein contacts.

* Corresponding author. Address: Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India. Fax: +91 80 23600551.
E-mail address: dpal@serc.iisc.ernet.in (D. Pal).

0014-5793/$36.00
doi:10.1016/j.febslet.2010.02.021

2. Materials and methods

2.1. Data sets

2.1.1. The complex set

Our data set contains atomic coordinates from the protein–protein interfaces covering area ranges between 240 and 7659 Å^2 available in the Protein Data Bank (PDB). Representative interfaces were derived from the quaternary structures of homodimeric and heterodimeric X-ray crystal structures using screening resolution and R-factor cut off of 2.5 Å and 0.2, respectively. Each subunit had at least 25 residues and no ligand of size five or more atoms at the interface. If any protein interface had an ambiguity in the atom-coordinates, we kept the one with the highest occupancy (electron density). Redundant complexes sharing >90% sequence identity of primary structure at the interfaces was excluded and only one representative chosen arbitrarily – this gave us a set of 906 unique protein dimers of which 800 were homodimers and the rest were heterodimers. Literature, curated PiQSi database [18] (version 2009-9-5), and the Protein Quaternary Structure (PQS) database [19] was used to verify that the interfaces chosen were biological. More stringent non-redundant subsets at 60%, 40% and 35% sequence identity at the interface contained 855, 640, 118 numbers of complexes, respectively.

2.1.2. The monomer set

A set of 386 monomers was chosen from the curated PiQSi database with crystal contact areas between 188 and 2111 Å^2, using resolution and R-factor cut off of 2.5 Å and 0.2, respectively. The monomers were non-redundant at <90% sequence identity.

2.1.3. The control set

Hundred protein complexes were arbitrarily chosen from the above *complex* set and atoms at the interface were deleted manually to decrease the surface complementarity and lessen the interface packing. The data set had artificial interfaces in the range of 344–2240 Å^2.

2.2. Interface area

We delineated the surface atoms of the protein molecule using the program NACCESS [http://wolf.bms.umist.ac.uk/naccess/] which calculates the solvent accessible surface area by rolling a probe atom of 1.4 Å radius over the surface of the protein [20]. Interface area was defined as the solvent accessible surface area lost/buried per subunit, on the complex formation.

2.3. Surface tessellation of proteins

Surface atoms of protein can be used to draw triangles that represent the surface geometry. This process of triangulation is called tessellation. A representative figure for tessellation of a protein surface is given in Fig. 1A. At first, an invariant frame of reference for the protein molecule is fixed for which the surface is to be tessellated. We calculate centroid of all the residue types present at the interface, as well as the centroid of all the interface atoms. The all-interface atom centroid is designated as the new origin. The *X*–*Y* plane of the new coordinate frame is determined by choosing two more centroids of different residues following a specific order that is non-collinear (distance >2.0 Å and angle between them >15°). These two centroids are used to assign the positive *X* and *Y* axis. The *Z*-axis is determined by the cross product of the two unit vectors along *X* and *Y* axis. The coordinates of the protein atoms are transformed to this new frame of reference before surface tessellation.

To tessellate the surface, we apply two-dimensional Delaunay triangulation [21,22] on very small sub-regions of the interface. All interface atoms are put into a grid of cell size 3.0 Å by the side. Delaunay tiling is applied on the interface atoms of each alternative grid cell along with its adjacent cell after projecting the atoms on a plane. See Flowchart 1 in Supplementary data for algorithm.

Because of the jagged nature of the rim region of the interface, some triangles formed may be uneven (smallest internal angle <30°). In order to minimize that, we iterate the tessellation procedure by changing the directionality along the axis. The best triangle from six such combination is picked from evenly formed triangles.

2.4. Surface complementarity

The surface of each subunit interface is tessellated as outlined above. Two triangles belonging to two different subunits across the interface are defined as *complementary* with each other if the distance between the centroids of the two triangles is $\leqslant$6.0 Å and the angle between the normal to the plane of the triangles $\leqslant$25°. Since the two complementary triangles might have disparate areas, we defined the triangle with minimum area among them as the complemented area. The triangles with side >5.0 Å and/or area >11.0 Å^2, mostly located at the interface rim are not considered for calculating complemented area. Total complemented area of an interface is calculated as the sum of all such complemented areas. The contribution of all the triangles on a subunit interface is defined as the total triangulated area of that subunit. If the interface area is small (<1000 Å^2) then triangles with side >5.0 Å and/or area >11.0 Å^2 are discarded from the triangulated set. Surface complementarity score is defined as the ratio of the total complemented area and the minimum of total triangulated area between the two subunits. Interface area correlated with surface complementarity with a correlation coefficient of −0.51. We divided surface complementarity score by interface area to obtain the normalized surface complementarity (NSc) measure. See Flowchart 2 in Supplementary data for algorithm.

2.5. Interface packing

An interface slice consisting of atoms within the 4.0 Å sphere radius of the interface atoms of each subunit of the complex is at first selected (Fig. 1B). The solvent accessible surface area of this slice using a probe radius 0.4 Å gives a surface area enveloping the inter-atomic void and the atoms. The enclosed volume (V_c) is taken as the volume of the sphere with equivalent solvent accessible surface area of the interface slice. Within the interface slice, the total volume (V_t) contributed by all the atoms can be estimated by considering each atom as a hard sphere with a specified van der Waals radius [23]. The ratio of V_t and V_c gives interface packing. Interface area correlated with interface packing with a correlation coefficient of −0.77. We divided interface packing by the interface area to get the normalized interface packing (NIP) measure.

2.6. Choice of threshold

The threshold values for NIP and NSc were chosen systematically in a combined manner. At first we evaluated NIP using probe radius of 0.2 Å, 0.4 Å, 0.6 Å, 0.8 Å, and 1.0 Å where we know the protein surface roughness estimated by fractal dimension D undergoes least fluctuation [24]. We then worked out NSc values for a combination of distance (5.0 Å, 6.0 Å, 7.0 Å) and angle (15°, 25°, 35°) thresholds. The Pearson correlation coefficients were calculated between the NIP and NSc values for the *complex* set. The minimum correlation for any combination of threshold parameters was found to be $\geqslant$0.9 (Supplementary Fig. S1); therefore, we decided on the 6.0 Å distance and 25° angle threshold for the NSc

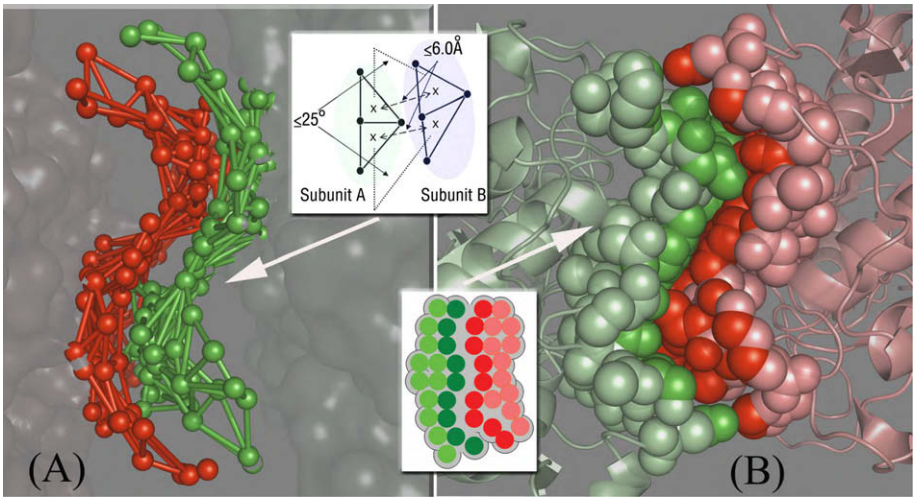

Fig. 1. Interface representation for calculation of NSc and NIP for the homodimeric protein complex PDB:2FUV, a phosphoglucomutase from *Salmonella typhimurium*. (A) Figure showing an example of tessellated surface at the protein–protein interface for calculation of NSc. Individual subunits are shown in a different color. In the inset, a schematic representation of the parameters used for NSc calculation are depicted. (B) Diagram showing the atomic slice selected for calculation of NIP. The contacting interface atoms are shown in a darker shade; whereas, the atoms within the 4 Å radius of the interface atoms are shown in a lighter shade of the same color as the interface atom. In the inset, we show a schematic representation of the interface slice.

calculation, which are neither too stringent nor too liberal. Similarly, we chose 0.4 Å as the probe radius for NIP calculations not to miss out on too many reentrant surfaces, yet not penetrate fractures or thin clefts that may be present at the interface. Furthermore, the minimum B-factor of atom-coordinates in Protein Data Bank (PDB) files is usually >10 $Å^2$, which yields a minimum coordinate-oscillation of about 0.4 Å around the equilibrium position using the Debye–Waller equation. Therefore, the solvent accessible surface area computed with probe radii of 0.4 Å is unlikely to err too much from the actual envelope volume of the interface slice. We have also varied the interface slice thickness before using the current threshold.

2.7. Availability

The program for computing NIP and NSc can be freely downloaded at http://pallab.serc.iisc.ernet.in/nip_nsc.

3. Results

We evaluated our measures along with three popular measures used for geometric complementarity for which softwares are freely available. (i) Shape complementarity index from Lawrence and Colman [15] as implemented in CCP4 suite (http://www.ccp4.ac.uk/html/sc.html) and normalized by interface area, hereafter called C-NSc, Gap Volume Index (GVI) from Laskowski [13], and Grid Correlation score based on Katchalski-katzir et al. [8] and normalized by interface area, and called NGC. Detailed results can be found from the Supplementary data.

3.1. Comparison of correlations on various datasets

3.1.1. The complex set

NIP and NSc values proposed by us show high correlation overall and across all interface area ranges (Table 1). When we compare NIP and NSc with other methods, only C-NSc shows an excellent overall correlation, and for all interface ranges. NGC shows a moderate overall correlation with NIP, NSc, and C-NSc, as well as for interface area $\leqslant$800 $Å^2$. GVI does not show any meaningful correlation overall and across any interface area range.

3.1.2. The monomer set

To further test on the efficiency of our measures, we took an independent set of crystal lattice contacts as available from the lattice contacts of monomers [18]. NIP, NSc, C-NSc and NGC values for each lattice contact correlated well among each other (Table 1, penultimate column, marked in bold). GVI did not show any correlation.

3.1.3. The control set

A consistent measure requires that it must also work well on control data set. The correlation between the NIP and NSc scores was obtained at 0.96 (Table 1, last column). NIP and NSc correlated very well with C-NSc, giving a correlation coefficient of 0.93 and 0.90, respectively. The same showed a somewhat lower correlation with NGC at 0.75 and 0.71, respectively. GVI did not show any useful correlation. This suggests that our method works equally well on surfaces that have a lower degree of packing and surface complementarity.

3.1.4. The non-redundant subsets

To ensure that the high correlation that we are getting for our methods is not for any bias in our data, we also evaluated the correlations on non-redundant subsets with 60%, 40% and 35% sequence identity at the interfaces. In all cases, we obtained correlation coefficient at 0.95.

3.2. Comparison of discriminative power of the measures

We also checked if the scores output by our method is useful in discriminating biological interfaces from crystal lattice contacts (which are non-biological). This discrimination assumes great significance when there are a number of protein–protein contacts in the crystal lattice that are of similar size and shape. The summary of the comparison between our methods and others are shown in Fig. 2. Among all the methods GVI fares weakly in discriminating crystal lattice contacts (*monomers*, white bars) from biological interfaces (*complexes*, black bars) and relatively less packed interfaces (*control*, gray bars) (Fig. 2A). C-NSc can segregate monomers from other interface types very well, especially at values >1.2 $\times$ 10^{-3} $Å^{-2}$ (Fig. 2B). Interestingly, NIP, which correlates very

Table 1
Correlation between various methods estimating surface complementarity/packing at the protein–protein interface.

Method		Pearson correlation coefficient					
		Complex				Monomer	Control
		Interface area range (Å^2)					
		All	⩽800	***800< and ⩽1500	>1500		
NIP	**NSc**	**0.95**	**0.79**	**0.85**	**0.91**	**0.88**	**0.96**
NIP	GVI	0.40	0.18	−0.08	0.16	0.07	0.39
NIP	**C-NSc**	**0.98**	**0.92**	**0.91**	**0.96**	**0.92**	**0.93**
NIP	**NGC**[a]	**0.66**	**0.70**	0.32	0.40	**0.83**	**0.75**
NSc	GVI	0.45	0.29	0.05	0.27	0.16	0.46
NSc	**C-NSc**	**0.95**	**0.75**	**0.81**	**0.93**	**0.83**	**0.90**
NSc	**NGC**[a]	0.65	**0.60**	0.28	0.45	**0.82**	**0.71**
GVI	C-NSc	0.44	0.25	−0.01	0.26	0.13	0.50
NGC[a]	GVI	0.26	0.33	−0.12	0.01	0.27	0.33
NGC[a]	C-NSc	**0.67**	**0.67**	0.33	0.44	**0.81**	**0.69**

Correlation coefficients ⩾0.6 are marked in bold. See Supplementary data for raw values on individual interfaces.
[a] Grid correlation score computed as by [7] in their FTDOCK program divided by interface area.

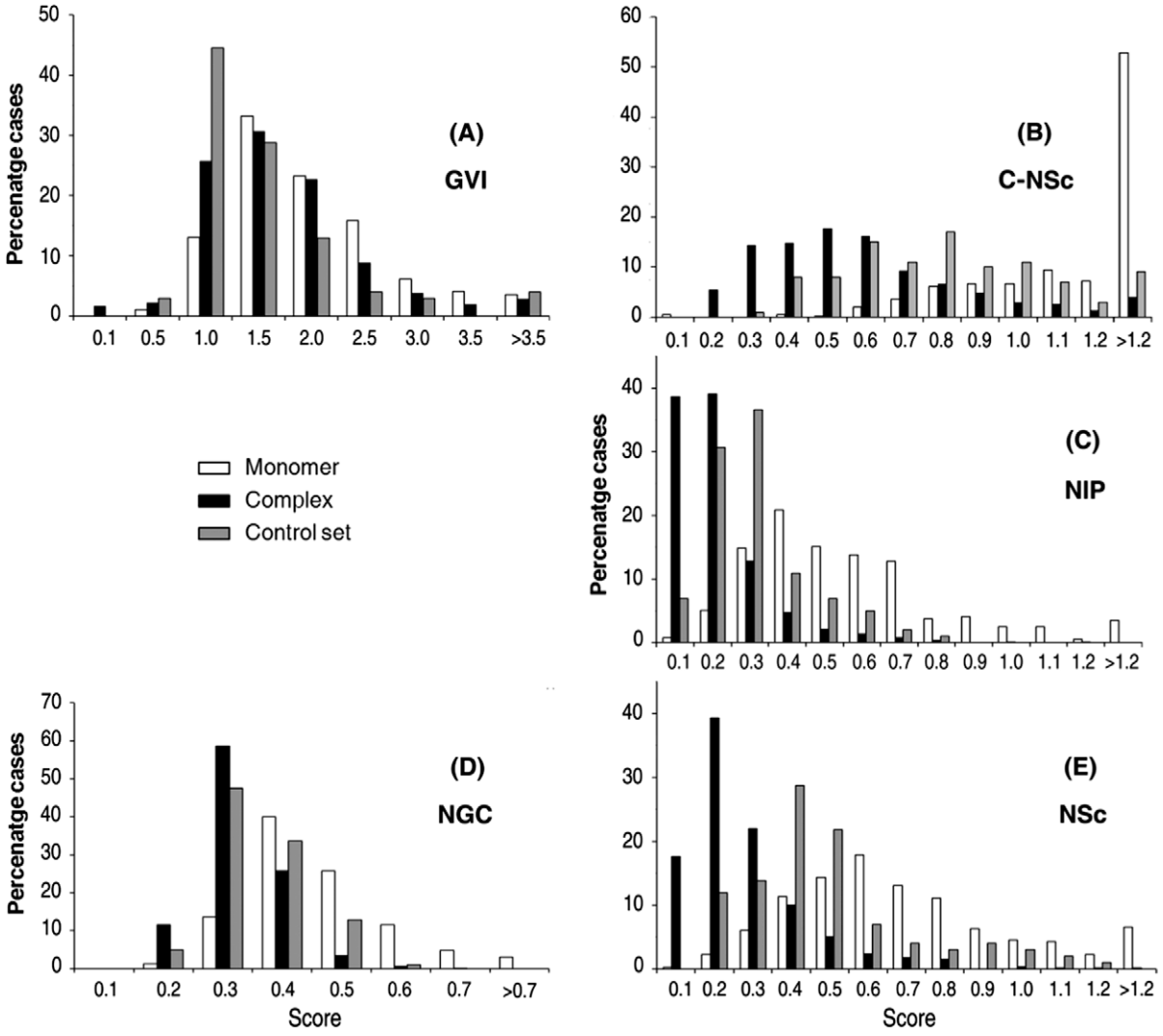

Fig. 2. Distribution of the output by individual methods on interfaces from different data sets: *Monomer*, *Complex* and *Control* set. Scores for individual methods are given in *X*-axis: (A). GVI (Å), (B) C-NSc (Å^{-2}), (C) NIP (Å^{-2}), (D) NGC (Å^{-2}), and (E) NSc (Å^{-2}). The C-NSc, NIP and NSc depicted in the plot are multiplied by 1000.

well with C-NSc, show efficient segregating power for biological interfaces, but not for monomers. NGC performs moderately well in discriminating biological interfaces from monomers (Fig. 2D), but better than GVI. NSc can separate crystal lattice contacts from biological interfaces more efficiently than NIP (Fig. 2E). At scores $<2 \times 10^{-4}$ Å^{-2} both NIP and NSc can almost exclusively identify a large proportion of the biological interfaces by excluding crystal lattice contacts.

Surveying an individual example, as in PDB:2FUV (to be published) (Fig. 1), where the surface complementarity values for the 1097 Å^2 interface are: NIP = 0.17, NSc = 0.23, GVI = 1.9, C-NSc = 0.72 and NGC = 0.40, one can see that only in our method the results are in a range (see Fig. 2 for relevant ranges) that is confidently separated from the Monomer set as well as the Control set values. This discriminative power is not well established for any of the existing methods. A visual inspection of the interface confirms the compact packing and complementarity of the atoms at the interface. One of the reasons we believe our method performs better is because we work exactly at van der Waals distance separation for estimation of atom packing and surface complementarity. In contrast, C-NSc uses the Connolly surface [17] which puts dots at uniform intervals smaller than the van der Waals separation of atoms. NGC calculation scheme has an implicit error since the method uses a 3D grid to estimate surface complementarity. As previously stated, GVI has fared unsatisfactorily in all our benchmarks, as it uses best-fitting spheres to estimate the void, which can leave a substantial amount of unaccounted interstitial space cumulated to erroneous results.

3.3. Performance

Our method is significantly more efficient in estimating NIP/NSc together in comparison to C-NSc (Fig. 3). This can be judged from the fact that to estimate C-NSc for a 7659 Å^2 interface in a 3.33 GHz workstation takes 1952 s of wall-clock time, in comparison to a mere 6.5 s for our NIP/NSc calculation. Grid correlation score (un-normalized version of NGC) is reported to have taken 9 s to compute for a homocomplex with 2200 atoms, in Convex C220 supercomputer with two CPUs (peak performance 50 MFLOPs each) [8]. Our performance using currently available workstations is definitely comparable (if not superior), which makes our method suitable for application in rapid surface complementarity and atom packing scans.

4. Discussion

Our methods are distinct in comparison to other methods in use for measuring surface complementarity and atom packing at the protein–protein interface. For the first time, we have used Delaunay triangulation for tessellation of protein surface atoms to measure surface complementarity. Delaunay's triangulation has been previously used on proteins in a different application for creation of simplices to evaluate four-body nearest neighbor propensities of amino acids [25]. A related tessellation technique, Voronoi polyhedra and alpha-complex have been used to measure atom packing, but the implementation is different. Walls and Sternberg [11] use alpha-complex to measure packing at the interface by first calculating volume for each atom type within the interface and comparing it with average volume occupied for each atom type on unrelated proteins. Li et al. [9] used surface dots as defined by the Voronoi surface projected into a cube/grid space and evaluated the surface normals of cubes across the interface to measure complementarity.

Our atom packing measure is also conceptually novel. We did not find any report in literature on the use of interface slice and envelope volume to compute atom packing at the interface.

Using two correlated measures, such as NIP and NSc in conjunction, offers an advantage that any classification made on the basis of these measures are likely to be more robust and offer increased discriminative power in assessing geometric properties at the protein–protein interfaces. It may be noted that because we deal exclusively with interface atoms irrespective of their atomic classification, our methods would work equally well for interfaces with/without water or other non-protein atoms. Combined with the time-efficient calculation of the quantities, offers a practical option and advantage of use in large-scale evaluation of surface complementarity and atom packing at the protein–protein interfaces.

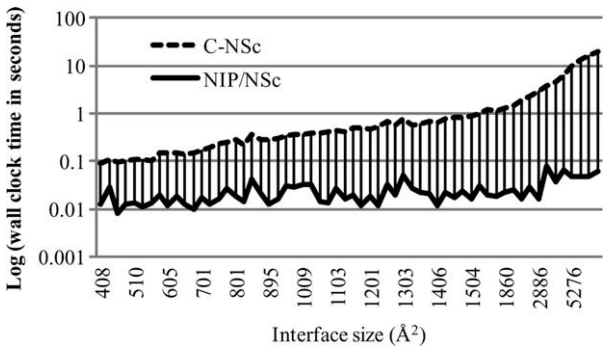

Fig. 3. Evaluation of computing–time performance between C-NSc and NIP/NSc in a 3.33 GHz processor workstation.

Acknowledgments

P.M. thanks All India Council for Technical Education, New Delhi for the national doctoral fellowship. The work was supported by funds from the Department of Biotechnology, New Delhi under the Centre of Excellence grants in Bioinformatics and Tuberculosis Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.02.021.

References

[1] Leicester, S.E., Finney, J. and Bywater, R. (1988) Description of molecular surface shape using Fourier descriptors. J. Mol. Graph. 6, 104–108.
[2] Max, N.L. and Getzoff, E.D. (1988) Spherical harmonic molecular surfaces. IEEE Comput. Graph. Appl. 8, 42–50.
[3] Ritchie, D.W. and Kemp, G.J. (2000) Protein docking using spherical polar Fourier correlations. Proteins 39, 178–194.
[4] Mitchell, J.C., Kerr, R. and Ten Eyck, L.F. (2001) Rapid atomic density methods for molecular shape characterization. J. Mol. Graph. Model. 19 (325–330), 388–390.
[5] Lo Conte, L., Chothia, C. and Janin, J. (1999) The atomic structure of protein–protein recognition sites. J. Mol. Biol. 285, 2177–2198.
[6] Chen, R., Li, L. and Weng, Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80–87.
[7] Gabb, H.A., Jackson, R.M. and Sternberg, M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. 272, 106–120.
[8] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA 89, 2195–2199.
[9] Li, N., Sun, Z. and Jiang, F. (2007) SOFTDOCK application to protein–protein interaction benchmark and CAPRI. Proteins 69, 801–808.
[10] Richards, F.M. (1977) Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6, 151–176.
[11] Walls, P.H. and Sternberg, M.J. (1992) New algorithm to model protein–protein recognition based on surface complementarity. Applications to antibody–antigen docking. J. Mol. Biol. 228, 277–297.
[12] Jones, S. and Thornton, J.M. (1996) Principles of protein–protein interactions. Proc. Natl. Acad. Sci. USA 93, 13–20.
[13] Laskowski, R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph. 13 (323–330), 307–308.
[14] Bahadur, R.P., Chakrabarti, P., Rodier, F. and Janin, J. (2004) A dissection of specific and non-specific protein–protein interfaces. J. Mol. Biol. 336, 943–955.
[15] Lawrence, M.C. and Colman, P.M. (1993) Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234, 946–950.

[16] Norel, R., Lin, S.L., Wolfson, H.J. and Nussinov, R. (1995) Molecular surface complementarity at protein–protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J. Mol. Biol. 252, 263–273.
[17] Connolly, M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science 221, 709–713.
[18] Levy, E.D. (2007) PiQSi: protein quaternary structure investigation. Structure 15, 1364–1367.
[19] Henrick, K. and Thornton, J.M. (1998) PQS: a protein quaternary structure file server. Trends Biochem. Sci. 23, 358–361.
[20] Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379–400.
[21] de Berg, M., van Kreveld, M., Overmars, M. and Schwarzkopf, O. (1997) Computational Geometry: Algorithms and Applications, Springer-Verlag, Berlin, Heidelberg.
[22] Guibas, L. and Stolfi, J. (1985) Primitives for the manipulation of general subdivisions and the computation. ACM Trans. Graphics. 4, 74–123.
[23] Lesk, A.M. and Chothia, C. (1980) Solvent accessibility, protein surfaces, and protein folding. Biophys. J. 32, 35–47.
[24] Lewis, M. and Rees, D.C. (1985) Fractal surfaces of proteins. Science 230, 1163–1165.
[25] Singh, R.K., Tropsha, A. and Vaisman, I.I. (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J. Comput. Biol. 3, 213–221.