Surface Lexography

Anthony Nicholls
Daylight Software Inc.
27401 Los Altos, Suite #370, Mission Viejo, CA 92691 USA
nicholls@daylight.com

1 Introduction

We have come to know and think about protein structures in terms of their various geometric representations (Fig. 1)
.

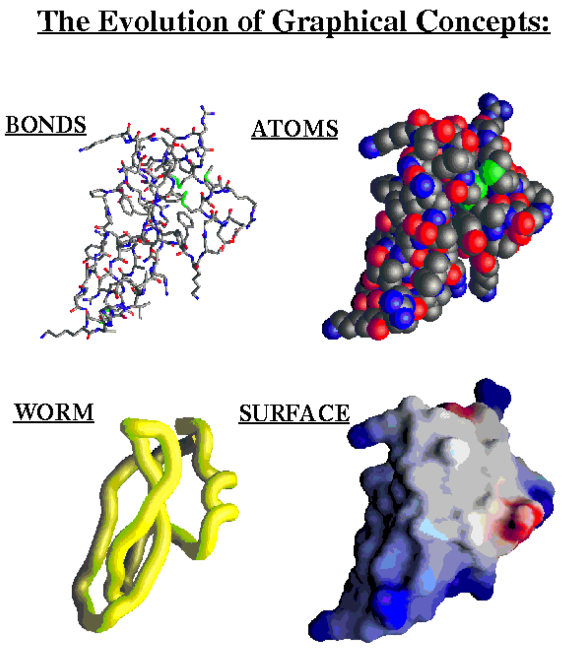


Fig. 1 Graphical Concepts

For instance, early molecular models were literally "built" with metal wires to represent bonds, or with plastic spheres for atoms. The age of the computer has brought perhaps less time consuming methods of building such models, as well as expanding our list of represention. These now include such constructs as backbone "worms" and molecular surfaces which, in addition to an increase in physical dimensionality over atoms and bonds, allows for new appreciation of the complexity of protein structure and structures relationship to function.

This historical process of structural conceptualization can be summarized as
i) visualization,
ii) pattern recognition and correlation with function,
iii) systematic annotation and query.
For instance, at the level of atomic positions the visualization of the common proximality of serine, histidine and aspartate in serine proteases was soon recognized as a pattern or "motif" of functional consequence. Such "atomic" patterns may now be searched for in a structural database. Similarly, the convergence of backbone traces into discrete families has become a major tool in the analysis of protein classification and function, e.g. the SCOP database.

This presentation will concentrate on the latest structural feature to emerge as having functional significance, namely surfaces. In particular I will focus on the aspects of electrostatic potential and surface curvature, although I will indicate how such an approach may be applied to other properties. The emphasis on electrostatics and curvature is due to the physical importance of the phenomena of solvation and hydrophobicity in protein interactions, and to their intimate connection to the boundary between protein and solvent. In the case of electrostatics, modern continuum theory assigns the dielectric discontinuity between wax-like proteins and high dielectric water to the molecular surface. Similarly the hydrophobic effect is typically correlated with exposed surface area and, in at least one theory, to the curvature of that surface.

Definition of some surfaces terms:

1) The molecular surface is that formed by the inner surface of a probe sphere (of radius to approximate a water molecule) as it rolls over a hard sphere representation of the molecule.

2) The accessible surface is the locus of the CENTER of that probe sphere. Note that each molecular surface is associated with exactly one point on the accessible surface, but that each accessible surface point may be associated with many molecular surface points.

3) The "curvature" of an accessible surface point will, in this paper, be taken to mean the fractional accessibility of a water molecule placed at this point compared to that of a water molecule placed against a flat plane. (Nicholls, Sharp and Honig, 1991)

4) The Van der Waals surface is that formed by those parts of each atomic sphere which lies outside of all other atomic spheres. Note that the VdW surface shares some points in common with the molecular surface, the difference in the latter being referred to as the "reenterant" surface.

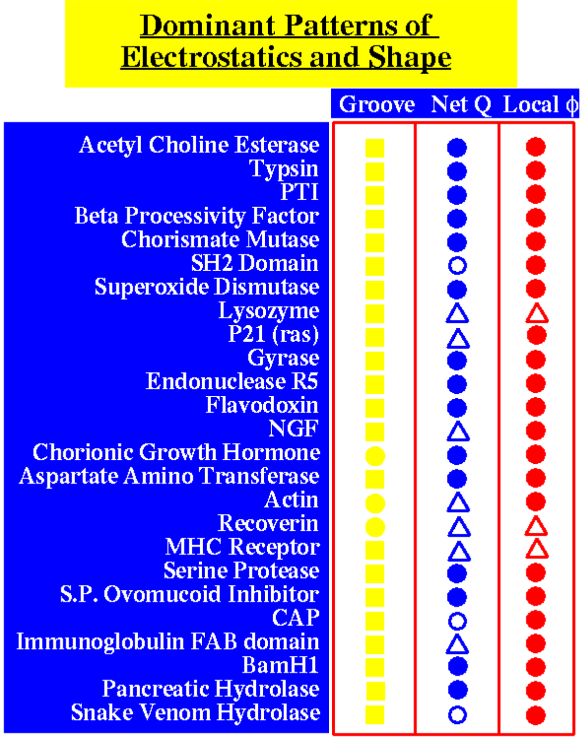
5) A contour surface is a general surface which separates 3D space into regions of greater than and less than a value associated with the surface. Note that molecular, accessible and VdW surfaces are all subclasses of contour surfaces.

6) Surfaces can be piecewise approximated by "tesselations", i.e. polygonal shapes joined at their edges to "cover" the surface. The tesselation more commonly used, and which I make use of herein, is the triangular tesselation.

7) A surface is termed closed if each triangle edge is shared with exactly one other triangle. If this number drops to zero for any edge the surface is termed open, or called a "patch". If this number rises to greater than one for any edge, the surface is ambiguous or "malformed".

Surface Patterns:

As the result of visualizing many disparate protein surfaces, I have noted the following general patterns (Fig 2):



Dominant Patterns of Electrostatics and Shape

	Groove	Net Q	Local ϕ
Acetyl Choline Esterase	■	●	●
Typsin	■	●	●
PTI	■	●	●
Beta Processivity Factor	■	●	●
Chorismate Mutase	■	●	●
SH2 Domain	■	○	●
Superoxide Dismutase	■	●	●
Lysozyme	■	△	△
P21 (ras)	■	△	●
Gyrase	■	●	●
Endonuclease R5	■	●	●
Flavodoxin	■	●	●
NGF	■	△	●
Chorionic Growth Hormone	●	●	●
Aspartate Amino Transferase	■	●	●
Actin	●	△	●
Recoverin	●	△	△
MHC Receptor	■	△	△
Serine Protease	■	●	●
S.P. Ovomucoid Inhibitor	■	●	●
CAP	■	○	●
Immunoglobulin FAB domain	■	△	●
BamH1	■	●	●
Pancreatic Hydrolase	■	●	●
Snake Venom Hydrolase	■	○	●

Fig. 2 General Patterns

1) The regions of proteins interaction are nearly always correlated with the largest, contiguously concave patch on the surface. Exceptions occur for weak ligands such as sugars and for ions.

2) Electrostatic complementarity between surfaces of protein and small ligand is rigorously enforced, although exceptions may occur within large protein-protein surfaces.

3) NET charge ANTI-complementarity is almost universal between proteins and small ligands.

4) Net charge complementarity IS seen however in proteins which bind to "macroscopic" charged substrates, e.g. charged membranes, dna, heparin.

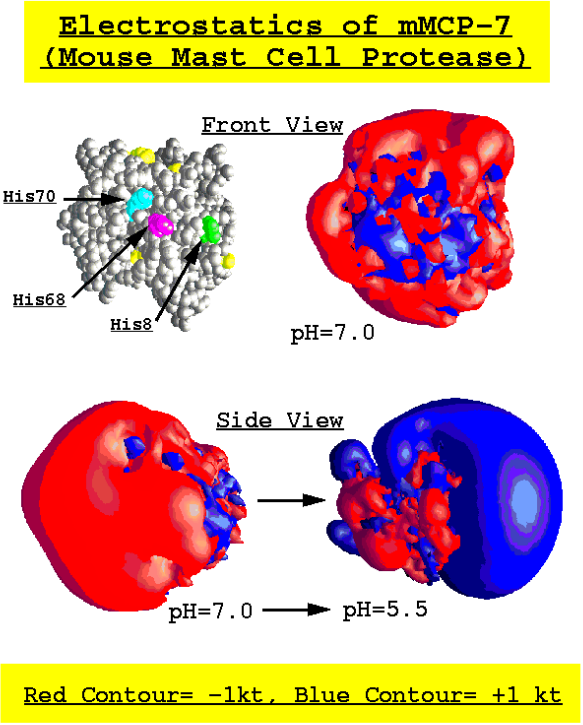
Examples of each pattern will be presented. The correlation of such patterns with protein function, and methods to make such observations automatic and quantifiable will constitute the bulk of this paper.

Predictive Ability:

The examples covered in this section will consist of work on Mast Cell Protease (Sali et al), Nerve Growth Factor (McDonald et al), human Chorionic Gonadatrophin (Wu et al), and HMG1 proteins (Landsman, Bryant and Baxevanis). Each example will utilize progressively greater quantification and automation of comparison procedures.

Mast Cell Protease:

Andrej Sali modeled a mast cell protease mMCP-7 using his program MODELLER and homology to the structure of Bovine Pancreatic Trypsin. The completed model showed an unusual distribution of surface histidines. The electrostatic contours (Fig. 3) at an energetically relevant level for charge states estimated at pH 7 and pH 5.5 showed a dramatic reversal of electrostatic character, from a predominantly negative protein to a predominantly positive protein.


Electrostatics of mMCP-7 (Mouse Mast Cell Protease)

Red Contour= -1kt, Blue Contour= +1 kt

Fig. 3 Electrostatics of mMCP-7

Furthermore, the location of the major changes in potential suggested three histidines are central to this polarity shift. Since mast cell proteases are hypothesised (Fig. 4) to associate with negatively charged heparin in secretory vacuoles kept at low pH, and to dissociate at neutral pH, it was hypothesized that these histidine where crucial to heparin binding. This was then confirmed via mutagenesis.

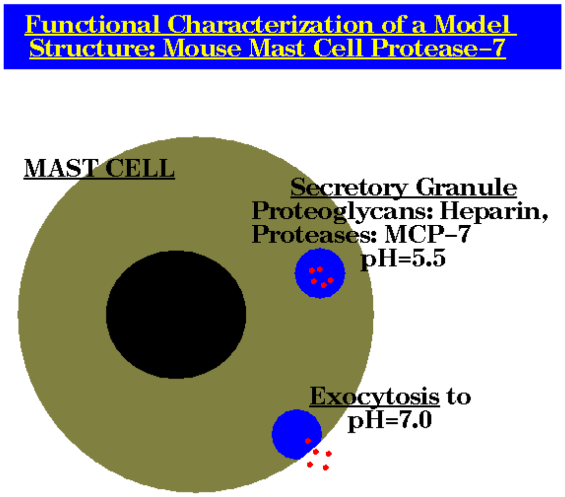

Functional Characterization of a Model Structure: Mouse Mast Cell Protease-7

Fig. 4 Mast Cell Model Structure

Nerve Growth Factor (NGF):

An unusual concentration of positive potential at one end of the dimer structure solved by McDonald et al was suggested and was found to correlate with low affinity binding behaviour by Ibanez et al. Murray-Rust (J.) noted that this site could also be predicted by modeling homologous trophic factors which bind to the same low affinity site, and comparing their surface electrostatics, i.e. the same positive patch was seen in each model, though other surface regions varied. McDonald and myself also have an outstanding prediction for the high affinity binding site based upon the largest contiguous concave patch on the surface of NGF.

Human Chorionic Gonadotrophin (CG):

Wu employed a similar graphical technique to Murray-Rust on her structure of hCG, modeling follicle stimulating hormone (FSH), thyroid stimulating hormone (TSH) and lutenizing hormone (LH). From these models a particular positive indentation can be seen which is close to a possible post-translational modification for the negatively charged sugar sialic acid.

Taking the structures Wu had generated I devised a simple method to reproduce the visual correlations she observed. First one defines a surface point to surface point map for each pair of aligned structures. In this case the mapping method is merely "minimal distance". Then the r.m.s. difference of the property value for each surface and each of its corresponding point on the other structural surface is calculated. This yields a new surface property, the homolog surface variability, in this case of electrostatic potential, though the method is not limited to such. Visual analysis leads to the conclusion that quantitatively the positive indentation observed visually is the largest conserved feature on the surface of the original structure. It also helps suggest where the receptor binding site might be located (by locating the most variable site). Combining this approach with that of locating the largest concave patches on the surface appears to be particularly powerful.

In addition to visual analysis, overall measures of surface similarity may be calculated from the values of surface variability normalized by total area. These measures confirm the visual perception that FHS appears to "cluster" with CG and LH with TSH.

High Mobility Group 1 (HMG1):

HMG1 proteins are a class of small nuclear binding proteins, with a distinctive "kinked" structure. The structure both with and without DNA has been solved by NMR (Werner et al) and crystallography (Read et al) respectively. Bryant, Landsman and Baxevanis modelled 84 homologous sequences on the crystal structure by a "threading" alignment procedure. Furthermore, they located 12 more sequences with no sequence homology but with a high "Z" score,indicative of a possible HMG1-like protein fold. Finally by combining all models they noted an overall average positive potential appeared in the then putative binding groove, consistent with DNA binding.

Applying the previous surface-surface metrics on so large a set of structures would have been very time consuming and so I constructed a second comparison metric (Fig. 5), this time more specific to electrostatics, which would be extremely rapid to construct. Working with Bryant's HMG1 homolog structures I calculated electrostatic "maps", i.e. potentials on a regular cubic mesh with encloses each structure and then found the r.m.s. potential difference between a structures surface potential as calculated from its own map from that calculated from a second structure map. This measure is then similar to the calculated for CG and homologs, but a hundred times faster.

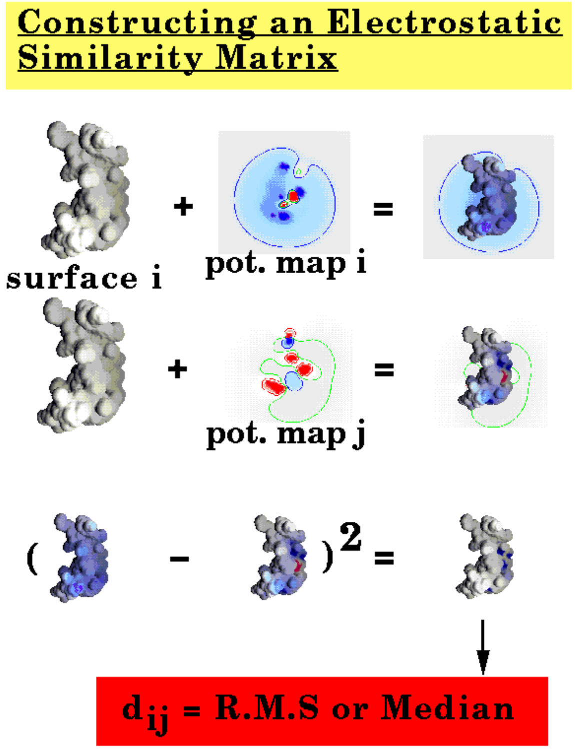
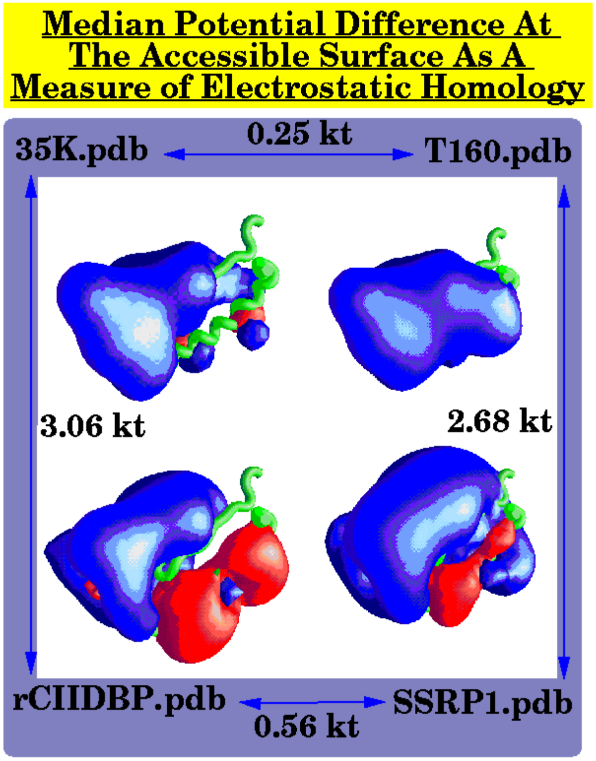

Constructing an Electrostatic Similarity Matrix
+
=
surface i
pot. map i
+
=
pot. map j
(
-
)
2
=
d_{ij} = R.M.S or Median

Fig. 5 Constructing an Electrostatic Similarity Matrix


Median Potential Difference At The Accessible Surface As A Measure of Electrostatic Homology
35K.pdb
0.25 kt
T160.pdb
3.06 kt
2.68 kt
rCIIDBP.pdb
0.56 kt
SSRP1.pdb

Thus all the surface-surface electrostatic homology scores (85*84/2) where calculated in a couple of hours on an SGI workstation. Visual analysis showed that indeed the similarity measure did appear to cluster analogous electrostatic patterns.

Given these scores a couple of analysis procedures proved useful. The first was to use these pair-wise scores in a clustering method, e.g. the Xcluster program in MacroModel developed by Shenkin at Columbia (Fig. 7).

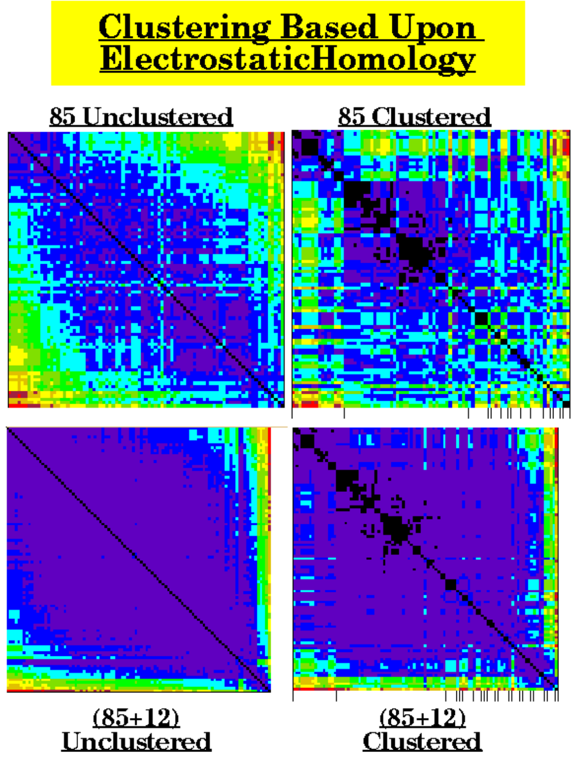


Fig. 7 Xcluster

This showed that the majority of structures fell into two classes (Fig.8). The first showed a monomodal potential distribution, the second a bimodal, where the later still has positive potential in the binding groove. This method also indicated that the 12 "novel" HMG1 structures did not appear to cluster with either group.

The second method was to take the surface of the original structure and determine if any parts of this surface were positive in ALL electrostatic maps of the 85 known homologs. This method, wherein any surface points which are negative from any single map are removed, or "digested", (Fig. 9) proved very enlightening. Four small surface patches remained after the procedure, three directly in the binding groove, two of such on convex projections, and the fourth nearby at the end of the protein. Furthermore, by loosening the selection criteria to allow at most ONE map to generate a positive potential, the entire inner binding groove of the protein "lit" up, along with an additional spot on the "elbow" of the protein.

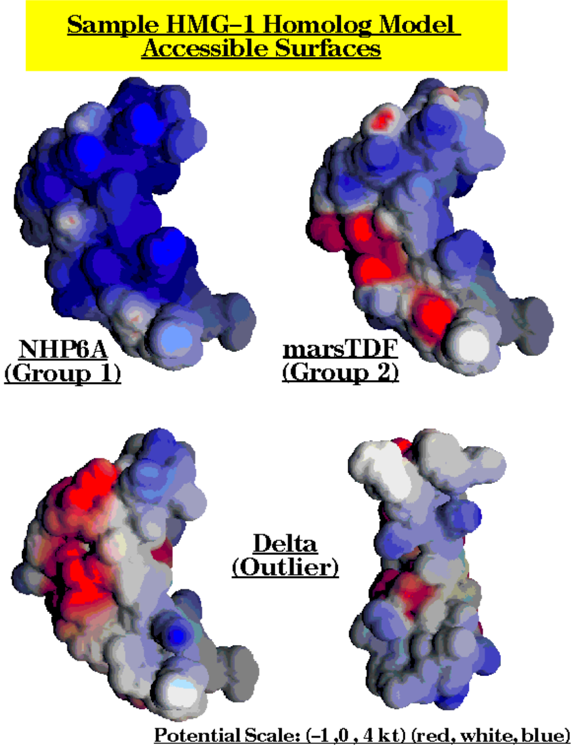

Sample HMG-1 Homolog Model
Accessible Surfaces
NHP6A
(Group 1)
marsTDF
(Group 2)
Delta
(Outlier)
Potential Scale: (-1 ,0 , 4 kt) (red, white, blue)

Fig. 8 Two Classes of HMG-1

These "consensus" surface patches could then be used to analyse the 12 putative HMG proteins, i.e. did their electrostatic maps agree with the consensus patches. The results of this analysis suggested that only two of the 12 appeared suitable HMG1 candidates, and with four giving completely opposite potentials in the critical regions.

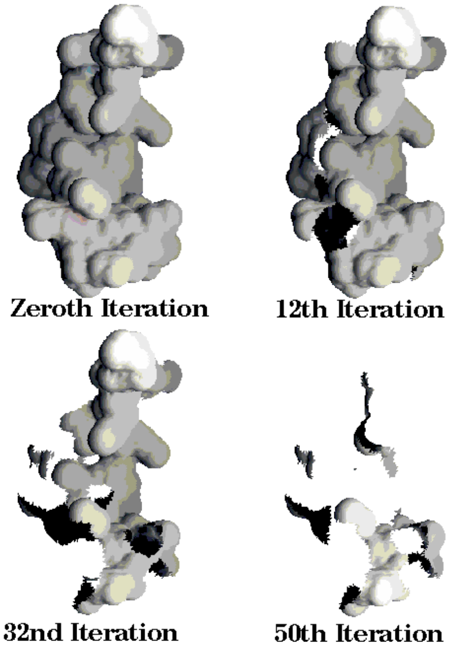
While it is currently unclear as to how "orthogonal" this surface information is to sequence information, it is clear that the application of electrostatic surface homology to a large set of models is potentially powerful.

Conclusions and Further Applications:

The concept of surface to surface comparison seems to hold considerable promise in analyzing models based upon sequence homology. Two methods of surface correlation where explored, surface-surface direct mapping and surface-field mapping. The latter is fast and suitable for electrostatics. The former, though slower, can be aggressively optimized and is more general in that any quantities can be compared (e.g.

hydrophobicity, mutability, chemical composition, curvature..). However, one major drawback at this time is that the alignment of structures relative to each other must be done relative to the underlying atoms. A more general method would dispense with the constituent atoms. Such methods are under development and promise to make surface "first class objects" suitable for database storage, retrieval and query.

Surface Digestion By Homologous Potential Maps (Positive Potential Condition)


Zeroth Iteration

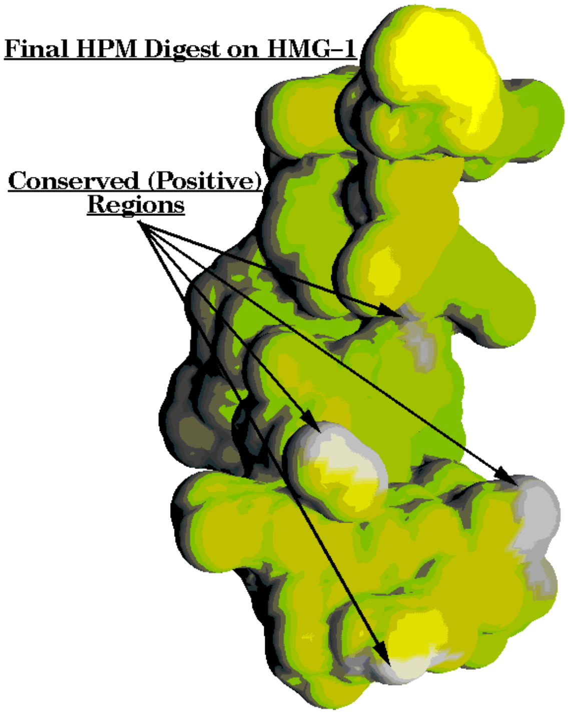
12th Iteration

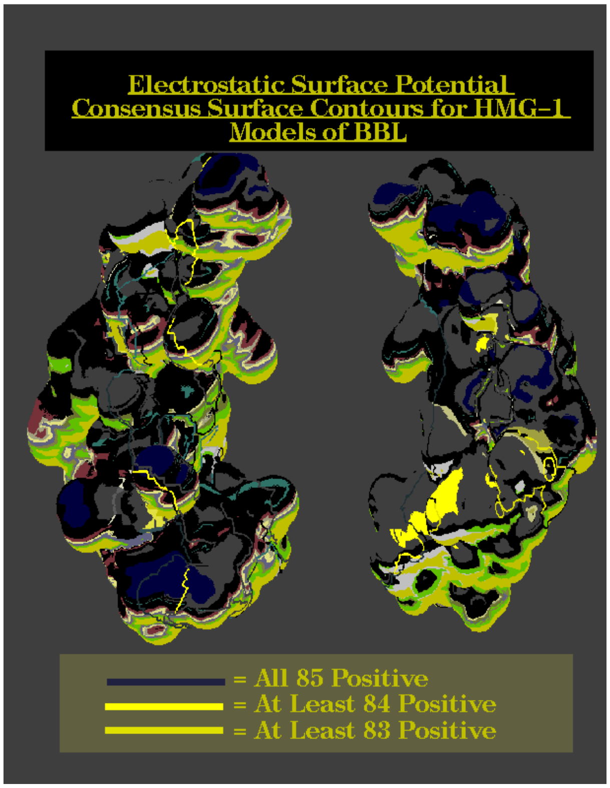
32nd Iteration

50th Iteration

.

Finally, these surface quantification concepts are equally applicable to complementary surfaces, e.g. of interacting proteins or domain-domain interactions within a single protein. For instance, the Surface Complementarity Index of Lawrence et al includes a similar surface-surface mapping function as utilized in the hCG example above. The application of surface of surface lexography techniques should also provide versatile and powerful in the characterization of protein- protein interactions.




Electrostatic Surface Potential Consensus Surface Contours for HMG-1 Models of BBL

Percent Homology Over Consensus Surface Region For the 12 "High Z Score" (Threading) / Low Sequence Homology Structures

	Q	% Homology
CHL1_YEAST	5.0	100
DMA_BPT2	1.0	14
DMA_BPT4	2.0	54
GCN1_YEAST	-1.0	26
MLE_DROME	5.0	100
MTF1_FLAOK	-7.0	0
PRRC_ECOLI	-5.0	0
RINA_BPPHA	0.0	27
RPA1_SCHPO	-5.0	19
U2AG_HUMAN	-11.0	0
VG2_BPLP7	1.0	36
YSCH_YERPS	-5.0	0

Measures of Surface Complementarity: Lawrence Surface Complementarity (LSC)

d_{ij} = distance between point i on surface A and its nearest point, j, on surface B

$$S^A(i) = (n_i \cdot n_j)\exp(-w d_{ij}^2)$$

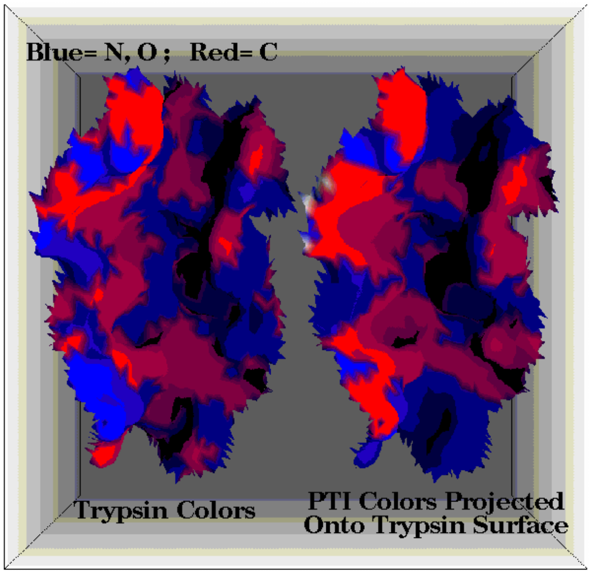
and vica versa for S^B.

LSC factor (S_c)

$$= 0.5 * (\text{median}\{S^A\} + \text{median}\{S^B\})$$

(Shape Complementarity at Protein/ Protein Interfaces, Michael C. Lawrence and Peter M. Colman, JMB, Vol. 234, pg. 946-950, 1993)

Hydrophobic/ Hydrophilic Complementarity: Trypsin/ PTI



Bibliography

Protein Folding and Association: Insights From the Interfacial and Thermodynamic Properties of Hydrocarbons. A. Nicholls, Kim Sharp and Barry Honig, PROTEINS, Vol. 11, 281-296 (1991)

Classical Electrostatics in Biology and Chemistry. Barry Honig and Anthony Nicholls, Science, 26th May, 1995, pp. 1144-1149 (1995)

The Electrostatic Basis for the Interfacial Binding of Secretory Phospholipases A2. David Scott, Arthur Mandel, Paul Sigler and Barry Honig, Biophysical Journal, Vol. 67, pp. 493-504 (1994)

An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif. Stephen Bryant and Chrarles Lawrence, PROTEINS: Vol. 16, pp92-112 (1993)

Homology Model Building of the HMG-1 Box Structural Domain. Andreas Baxevanis, Stephen Bryant and David Landsman

Structure of Human Chorionic Gonadotrophin at 2.6A Resolution from MAD analysis of the Selenomethionyl Protein. HaoWu, Joyce Lustbader, Yee Liu, Robert

Canfield andWayne Hendrickson, Structure, Vol 2, No. 6, pp 545-558 (1994)

A New Fold Revealed by a 2.3A Resolution Crystal Structure of Nerve Growth Factor. N. Q. McDonald, R. Lapatto, J. Murray-Rust, J. Gunning, A. Wlodawer and T.L. Blundell, Nature Vol 354, pg. 411

Three-dimensional models of four mouse mast cell chymases: Identification of proteoglycan-binding regions and protease-specific antigenic epitopes. Andrej Sali, R. Matsumoto, H.P. McNeil, M. Karplus and R.L. Stevens, J. Biol. Chem., pp. 9023-9034, (1993)

Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells: A cluster of histidines in mouse mast cell protease-7 regulates its binding to heparin serglycin proteoglycan. R. Matsumoto, A. Sali, N. Ghildyal, M. Karplus and R.L. Stevens.

Shape Complementarity at Protein/ Protein Interfaces, Michael C. Lawrence and Peter M. Colman, Vol 234, pg. 946-950, 1993) lexography techniques should also provide versatile and powerful in the characterization of protein- protein interactions.