1. DEVELOPMENT OF THE AESOP COMPUTATIONAL FRAMEWORK FOR ELECTROSTATICS-BASED PROTEIN DESIGN

1.1. ELECTROSTATIC CALCULATIONS IN PROTEIN INTERACTIONS

1.1.1. Two-step model of association

For studies focused on electrostatically-driven protein association, McCammon and coworkers have proposed a two-step association model that is used to deconstruct association, as well as the interactions that drive it [1]. In this two-step association model, the first step is known as recognition and consists of the initial collision of the two proteins free in solution through diffusive motion. Recognition is driven and or accelerated by long-range electrostatic interactions, and results in a weak non-specific encounter complex. This is then subsequently followed by the binding step, where short to medium range electrostatic interactions, van der Waals interactions, as well as entropic effects, drive the formation of a specific final complex. This model holds true for interactions between highly charged proteins and ligands, and is essential in understanding why mutations away from the binding interface can affect binding. An illustration of the two-step binding models is presented in Figure 1-1A. According to this model, long range electrostatic interactions are vital to the recognition or association phase of binding, and therefore, mutations that alter the spatial distribution of electrostatic potential can affect binding, even if the residue is away from the binding interface. This logic is different from traditional thinking, and is evidence for the need for a thorough electrostatic analysis.

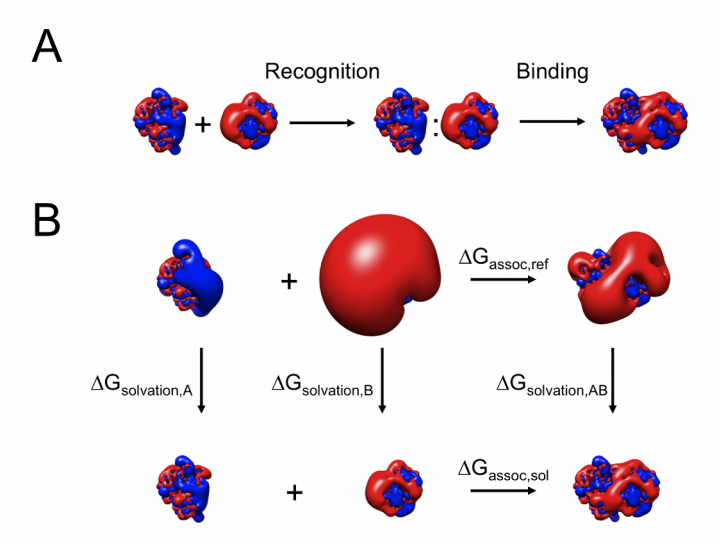

Figure 1-1 *Illustrations of the decomposition of the role of electrostatics in protein association.* (A) Schematic of the two-step model of association illustrated using isopotential contours of barnase-barstar. (B) Thermodynamic cycle for calculating solvation free energies of association illustrated using isopotential contours of barnase-barstar. The top process corresponds to association in the reference state and bottom process corresponds to association in the solution state (described in text). The color scheme of the isopotential contours is red for negative potential, and blue for positive potential.

1.1.2. Electrostatic similarity calculations

Based on the two-step model for association, it is intuitive that proteins with similar electrostatic properties, or more specifically similar spatial distributions of electrostatic potential, are likely functional homologues. Methods for quantifying electrostatic similarity have thus been of great interest for fields such as protein engineering and drug design. Early work in this field was centered in pharmacochemistry, where molecular similarities were used to compare the electron densities of small organic compounds based on quantum mechanical calculations. Carbo and coworkers [2] developed a similarity index (SI) for comparing molecular density functions, which has the advantage of being firmly grounded in quantum mechanics,

$$SI_{Carbo} = \frac{\int_V \rho_A \rho_B \, dV}{\left(\int_V \rho_A^2 dV\right)^{1/2} \left(\int_V \rho_B^2 dV\right)^{1/2}} \qquad \text{Eq. 1-1.}$$

Where, ρ_A and ρ_B represent density functions of the two molecules to be compared. Hodgkins and coworkers [3] were first to propose the use of a similarity measure for the comparison of molecular electrostatic potential or electric field. The Hodgkin electrostatic similarity index (ESI), is based on a dot-product and produces values from -1 to 1,

$$ESI_{Hodgkin} = \frac{2\sum \phi_A(i,j,k)\phi_B(i,j,k)}{\sum \phi_A(i,j,k)^2 + \sum \phi_B(i,j,k)^2} \qquad \text{Eq. 1-2.}$$

Where a value of 1 indicates identity, a value of -1 indicates anticorrelation, and 0 indicates no similarity. In this expression, the density functions have been replaced with the electrostatic potentials, ϕ_A and ϕ_B, and normalization is achieved using the sum of the self dot-products. Summations over all grid points *(i,j,k)* are performed, since space is discretized for numerical calculations of electrostatic potential. Two alternative measures, one by Reynolds and coworkers [4] and another by Petke [5], have also been proposed, both of which provide a linear relationship with respect to the proportionality of the compared electrostatic potentials,

$$ESI_{Reynolds} = \left[1 - \frac{1}{N}\sum_N \frac{\left|\phi_A(i,j,k) - \phi_B(i,j,k)\right|}{\max\left(\left|\phi_A(i,j,k)\right|,\left|\phi_B(i,j,k)\right|\right)}\right] \qquad \text{Eq. 1-3,}$$

$$ESI_{Petke} = \frac{1}{N}\sum_N \frac{2\,\phi_A(i,j,k)\,\phi_B(i,j,k)}{\phi_A(i,j,k)^2 + \phi_B(i,j,k)^2} \qquad \text{Eq. 1-4.}$$

Both the Reynolds and Petke ESIs utilize an average similarity that is locally normalized at each grid point. All three of these early ESIs (Hodgkins, Reynolds, and Petke) compared only potential values outside the van der Waals surface of the molecules. Wade and coworkers have since extended the ESI originally proposed by Hodgkins et al to the analysis of protein interactions [6,7]. However, Wade et al introduced the concept of a “skin” region, or a thin region of chosen thickness surrounding the protein, to account for the intricacies that arise when applying ESI methods to proteins. The skin region is used to focus the comparison to regions of functional importance, and to exclude large potential values that

arise in the protein interior. ESI values can also be converted into electrostatic similarity distances (ESD), which allow for the application of clustering algorithms. ESD measures simply require that values of zero indicate identity, whereas increasing values indicate increasing dissimilarity. Three examples of ESD measures derived from Eq. 1-2, Eq. 1-3, and Eq. 1-4, are as follows:

$$DP = \sqrt{1 - \frac{2\sum \phi_A(i,j,k)\phi_B(i,j,k)}{\sum \phi_A(i,j,k)^2 + \sum \phi_B(i,j,k)^2}} \qquad \text{Eq. 1-5,}$$

$$LD = \frac{1}{N}\sum_N \frac{\left|\phi_A(i,j,k) - \phi_B(i,j,k)\right|}{\max\left(\left|\phi_A(i,j,k)\right|, \left|\phi_B(i,j,k)\right|\right)} \qquad \text{Eq. 1-6,}$$

$$LDP = \sqrt{1 - \frac{1}{N}\sum_N \frac{2\,\phi_A(i,j,k)\,\phi_B(i,j,k)}{\phi_A(i,j,k)^2 + \phi_B(i,j,k)^2}} \qquad \text{Eq. 1-7.}$$

1.1.3. Electrostatic free energies of association

To quantify the electrostatic contributions to protein association, electrostatic free energies of association can ultimately be calculated, and such free energies have been shown to serve as good predictors of binding ability for highly and oppositely charged proteins [8-10]. When calculating electrostatic free energies of association, it is often of interest to incorporate solvation and other effects using a thermodynamic cycle such as that in Figure 1-1B. Solvation free energies of association, $\Delta\Delta G_{solv}$, which account for both solvation and association, can be calculated according to this thermodynamic cycle and the following expression,

$$\begin{aligned} \Delta\Delta G_{solv} &= \Delta G^{AB}_{solvation} - \Delta G^{A}_{solvation} - \Delta G^{B}_{solvation} \\ &= \Delta G_{assoc,sol} - \Delta G_{assoc,ref} \end{aligned} \qquad \text{Eq. 1-8.}$$

This thermodynamic cycle accounts for association in a uniform dielectric reference state, $\Delta G_{assoc,ref}$, without the presence of counterions, as well as a solvated state, $\Delta G_{assoc,sol}$. Additionally, the vertical processes represent solvation, $\Delta G_{solvation,(A,B,or\ AB)}$, of the two free components, as well as the complex, and aid in removing grid artifacts as has been discussed extensively [8].

1.2. DESIGN OF THE AESOP FRAMEWORK

1.2.1. Motivation

Based on the two-step association model (Figure 1-1A), mutations of ionizable residues away from the binding interface can affect protein association, by altering recognition. Therefore, from the point of view of protein design, mutations of ionizable residues provide an interesting advantage, since they can affect both the overall protein electrostatic potential (global), as well as more specific intermolecular interactions (local). We desired to create a method, based on the perturbative design approach, which could systematically evaluate the role of each ionizable residue in protein association and stability. The goal was to develop a framework for identifying electrostatic "hot spots" and optimizable sites, in order to aid in the design of new protein analogs with customized electrostatic properties. The resulting computational framework is referred as Analysis of Electrostatic Similarities of Proteins (AESOP) [9,11], and utilizes theoretical mutations, Poisson-Boltzmann electrostatics, and electrostatic similarity clustering to evaluate the role of electrostatics in protein association.

1.2.2. Computational workflow of AESOP

The general workflow of the AESOP framework, as illustrated by Figure 1-2, starts with a protein complex and the generation of electrostatically perturbed protein analogs. The standard analysis involves the use of theoretical alanine-scan mutagenesis in which each ionizable (charged) residue is replaced by alanine, one at a time. Additional perturbation methods, such as charge reversals and mutation permutations, are also included in the AESOP framework. Following the generation of perturbed mutant structures, there are two types of electrostatic calculations that are used to quantify the effects of the perturbations: electrostatic similarity clustering and electrostatic free energies of association. Electrostatic similarity clustering depicts the global effects of perturbations and relates to

the recognition step of association, while electrostatic free energies of association capture both local and global effects, and therefore correspond to both recognition and binding.

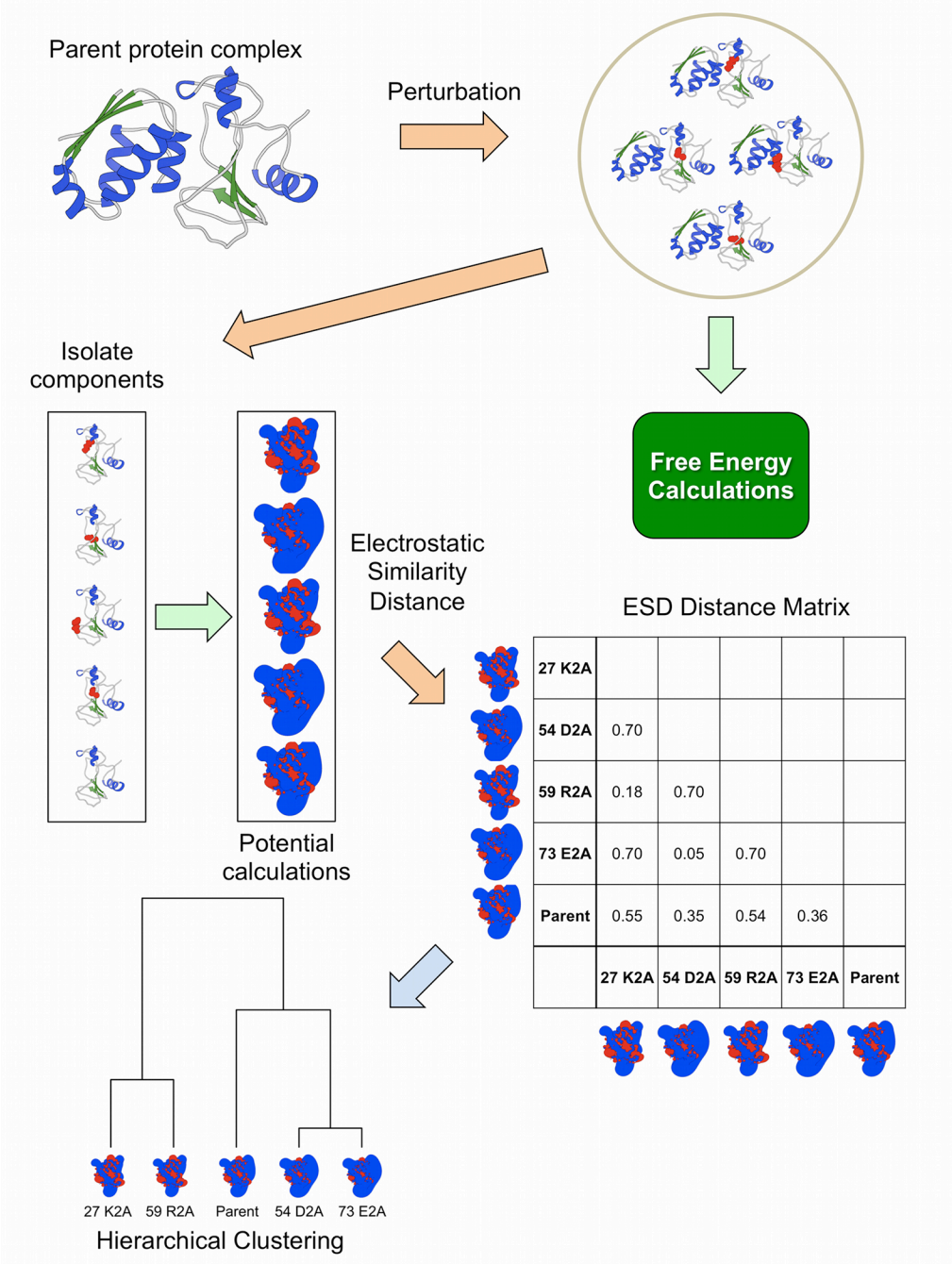

Figure 1-2 *Workflow of the AESOP Framework.* Arrow color indicates which utility is responsible for performing the various steps: orange, AESOP; green, APBS [13]; blue, R functions [15]. In ribbon model, red residues indicate perturbation site.

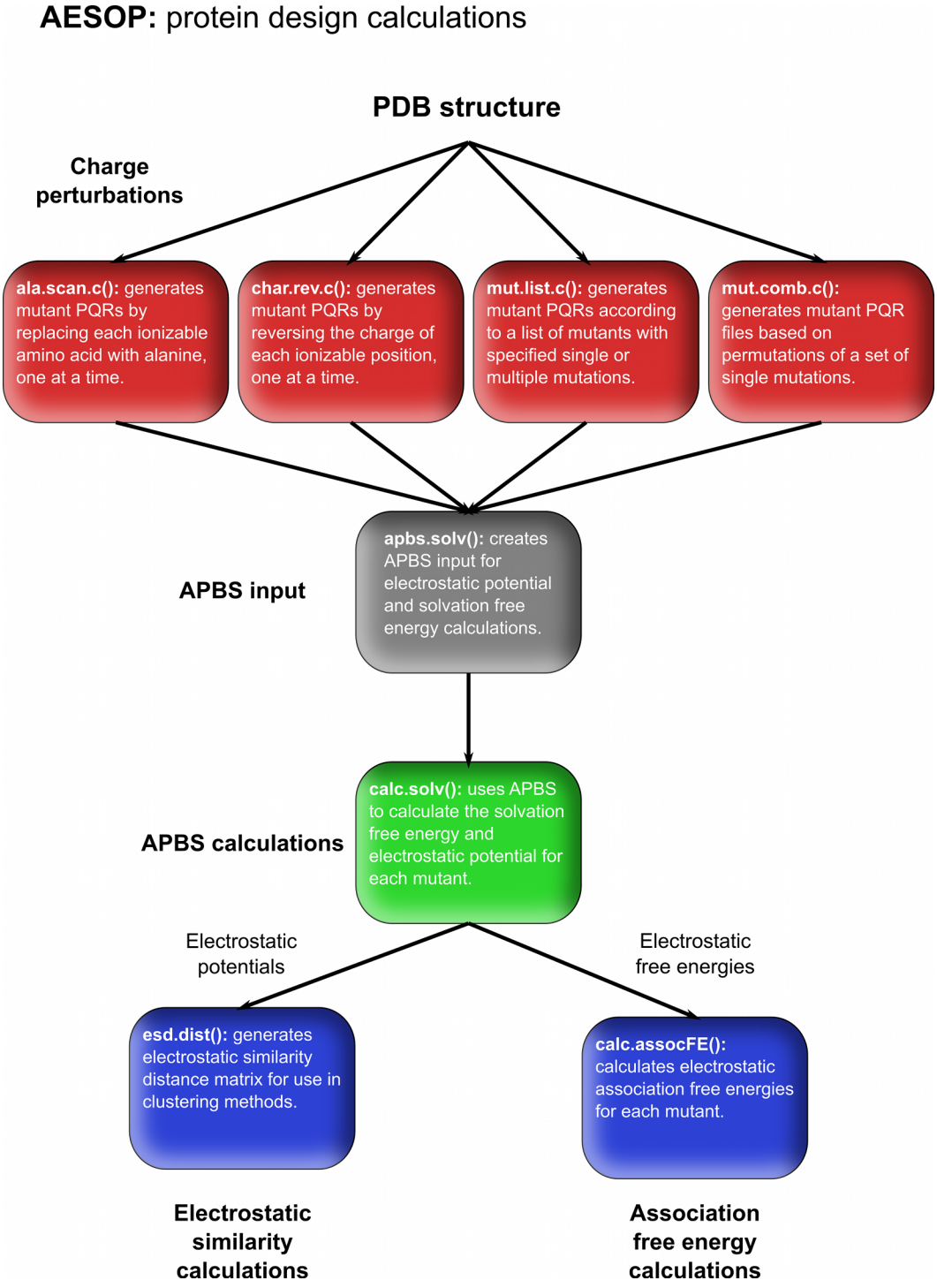

Figure 1-3 *Schematic of AESOP functions for electrostatics-based protein design.*

To perform electrostatic similarity clustering, the mutated proteins are first isolated from the complex, and the spatial distributions of each mutant are calculated using the Poisson-Boltzmann method. A distance matrix is then populated by ESD values, according to Eq. 1-5, Eq. 1-6, or Eq. 1-7, quantifying the similarity between spatial distributions of electrostatic potential for each possible pair of mutants. Clustering methods are applied to the ESD distance matrix to classify the perturbations according to their effects on the spatial distributions of electrostatic potential; one such method, is hierarchical clustering [12], which produces a dendrogram tree as shown on the bottom of Figure 1-2. Electrostatic free energies of association are typically calculated based on a thermodynamic cycle that accounts for solvation effects by including association in both a solvated state as well as a reference state (Figure 1-1B). All electrostatic calculations in the AESOP framework (green arrows, Figure 1-2) are performed using the Adaptive Poisson-Boltzmann Solver (APBS) [13]. It should also be noted that the AESOP protocol could be applied in the absence of a protein complex, to families of either homologous proteins or alanine-scan mutants. In general, the AESOP was developed with flexibility in mind in order to allow for customized electrostatic analyses.

1.2.3. AESOP library structure and usage

The AESOP framework has been written using the R statistical language [15] and utilizes the Bio3D [16] library to handle PDB structure data. The AESOP framework relies on three external software packages: (i) SCWRL4 [17] for non-alanine mutations, (ii) PDB2PQR [14] for structure preparation and parameterization, and (iii) APBS [13] for all electrostatic potential and free energy calculations. As illustrated by Figure 1-2, all AESOP calculations begin with a PDB file which is read into the R environment using the read.pdb() function of Bio3D, which allows usage of Bio3D functions when making complex atom/residue selections and other manipulations based on a PDB file. In the electrostatic-based design platform (Figure 1-3), the first primary step is the generation of electrostatically perturbed protein mutants. The AESOP framework includes 4 methods for introducing

charge perturbations: ala.scan.c(), char.rev.c(), mut.list.c(), and mut.comb.c(). ala.scan.c() performs a theoretical alanine-scan by replacing each charged residue with alanine, one at a time. These alanine mutations are performed by an AESOP function called mut2ala(), which simply truncates the residue down to the C_β atom and generates the appropriate hydrogen geometry. In addition to generating a directory containing the alanine-scan mutants, ala.scan.c() also returns a list of the mutants generated and the directories to which they were written. char.rev.c() is similar to ala.scan.c, however, instead of alanine mutations, the charge of each charged site is reversed one at a time by replacing every basic residue with glutamic acid and every acid residue with lysine. The charge reversal mutations are introduced using SCWRL4, which optimizes the side chain rotamer upon mutation. The remaining two functions are somewhat different in that only specified positions are mutated, not every charged residue as in ala.scan.c() and char.rev.c(). The mut.list.c() function takes as input a list of specific mutants, which can contain single or multiple mutations, and generates the mutated PQR files using a combination of SCWRL4 and PDB2PQR. In contrast, mut.comb.c() accepts a list of single mutations and then generates PQR files for all permutations, using SCWRL4 and PDB2PQR, given a desired number of mutations per mutant. Caution should be used when using the mut.comb.c() approach, since the number of permutations can become extremely large for a relatively small list of single mutations depending on the number of mutations per mutant.

After generating PQR files for charge-perturbed structures, the next step is to setup the APBS parameters for the electrostatic potential and free energy calculations. The function apbs.solv() generates an object containing the APBS parameters based on the initial PDB and the location of perturbed PQR files. apbs.solv() will assign default values to parameters such as protein dielectric and will suggest grid lengths based on the coordinates of the protein, however, all parameters can be adjusted using the APBS keyword notation. It is imperative at this step that the parent structure is used in centering the electrostatic calculations to remove the possibility of grid artifacts. Once the APBS input

parameters are initialized, the next step is to perform the Poisson-Boltzmann electrostatic calculations. For efficiency, the calc.solv() function simultaneously calculates the solvation free energy and electrostatics potential of each mutated component/complex. Instead of calculating the six states of the thermodynamic cycle (Figure 1-1B), followed by an additional calculation for the electrostatic potential, calc.solv() calculates the vertical process of for the parent and each mutant and saves the electrostatic potential for solvated state.

The final steps of the AESOP protein design platform involve quantifying/comparing the effects of the perturbations on the electrostatic character. The first approach comparison is electrostatic similarity clustering, which requires the esd.dist() function to generate an electrostatic similarity distance matrix, containing all pair-wise comparisons. The esd.dist() function simply requires the name of the directory containing the electrostatic potentials to be compared, and returns a two-dimensional distance matrix that can used in clustering methods included in the R statistical language, such as hclust() for hierarchical clustering. The second step in quantifying the effects of the electrostatic perturbations is to calculate solvation free energies of association by combing the free energies of the three vertical processes according to Eq. 1-8.

2. REFERENCES

1. McCammon J, Northrup S, Allison S (1986) Diffusional dynamics of ligand receptor association. J Phys Chem-Us 90: 3901–3905.
2. Carbo R, Leyda L, Arnau M (1980) How similar is a molecule to another - An electron-density measure of similarity between 2 molecular-structures. Int J Quantum Chem 17: 1185–1189.
3. Hodgkin E, Richards W (1987) Molecular similarity based on electrostatic potential and electric-field. Int J Quantum Chem: 105–110.
4. Reynolds C, Burt C, Richards W (1992) A linear molecular similarity index. Quant Struct-Act Rel 11: 34–35.
5. Petke J (1993) Cumulative and discrete similarity analysis of electrostatic potentials and fields. J Comput Chem 14: 928–933.
6. Wade R, Gabdoulline R, De Rienzo F (2001) Protein interaction property similarity analysis. Int J Quantum Chem 83: 122–127.
7. Blomberg N, Gabdoulline R, Nilges M, Wade R (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins 37: 379–387.
8. Cheung AS, Kieslich CA, Yang J, Morikis D (2010) Solvation Effects in Calculated Electrostatic Association Free Energies for the C3d-CR2 Complex and Comparison with Experimental Data. Biopolymers 93: 509–519. doi:10.1002/bip.21388.
9. Kieslich CA, Gorham RD, Morikis D (2011) Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. J Non-Cryst Solids 357: 707–716. doi:10.1016/j.jnoncrysol.2010.05.087.
10. Gorham RD, Kieslich CA, Nichols A, Sausman NU, Foronda M, et al. (2011) An evaluation of Poisson-Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. Biopolymers 95: 746–754. doi:10.1002/bip.21644.
11. Kieslich CA, Morikis D, Yang J, Gunopulos D (2011) Automated computational framework for the analysis of electrostatic similarities of proteins. Biotechnol Progr 27: 316–325. doi:10.1002/btpr.541.
12. Jain A, Murty M, Flynn P (1999) Data clustering: A review. Acm Comput Surv 31: 264–323.
13. Baker N, Sept D, Joseph S, Holst M, McCammon J (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. P Natl Acad Sci Usa 98: 10037–10041.
14. Dolinsky T, Nielsen J, McCammon J, Baker N (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 32: W665–W667. doi:10.1093/nar/gkh381.
15. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. pp. Available:http://www.R-project.org/.
16. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22: 2695–2696. doi:10.1093/bioinformatics/btl461.
17. Krivov GG, Shapovalov MV, Dunbrack RLJ (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77: 778–795. doi:10.1002/prot.22488.

3. AESOP APPLICATIONS

Reviewed in:
Gorham, R., Kieslich, C. A., and Morikis, D. (2011) Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Ann. Biomed Eng,* 39(4): 1252-63.

Exploration of effects of electrostatics parameters on comparison with experiment:
Gorham, R.D., Kieslich, C.A., Nichols, A., Sausman, N.U., Foronda, M., and Morikis, D. (2011) Calculation of free energy of protein association using Poisson-Boltzmann electrostatics: validation with experimental kinetic data. *Biopolymers,* 95(11): 746-54.

Comparison of electrostatic similarity measures:
Kieslich, C.A., Gorham, R.D., and Morikis, D. (2011) Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. *J. Non-Cryst. Solids,* 357(2): 707-716.

Hakkoymaz, H., Kieslich, C.A., Gunopulos, D., and Morikis, D. (2011) Molecular similarity determination using multi-resolution analysis. *Mol. Inform.,* 30(8): 733-46.

Various of Applications:

1. Kieslich, C.A. and Morikis, D. (2012) The two sides of complement C3d: Evolution of electrostatics in a link between innate and adaptive immunity. *PLoS Comp. Bio.* Submitted
2. Bellows-Peterson, M.L., Fung, H.K., and Floudas, C.A., Kieslich, C.A., Zhang, L, and Morikis, D, Wareham, K.J., and Monk, P.N., Hawksworth, O.A., and Woodruff, T.M. (2012) De novo peptide design with C3a receptor agonist and antagonist activities: Theoretical predictions and experimental validation. *J. Med. Chem.,* 55(9): 4159-4168
3. El-Assaad, A.M., Kieslich, C.A., Gorham, R.D., and Morikis, D. (2011) Electrostatic exploration of the C3d-FH4 interaction using a computational alanine scan. *Mol. Immunol.*, 48(15/16): 1844-50.
4. López de Victoria, A., Kieslich, C.A., Rizos, A.K., Krambovitis, E., and Morikis, D. (2011) Clustering of HIV-1 subtypes based on gp120 V3 loop electrostatic properties. *BMC Biophys.,* 5(3):1-16.
5. Gorham, R.D., Kieslich, C.A., and Morikis, D. (2011) Complement Inhibition by Staphylococcus aureus: Electrostatics of C3d–EfbC and C3d–Ehp Association. *Cell. Mol. Bioeng.,* 48(1): 32-43.
6. Kieslich, C.A., Goodman, G., Vazquez, H., López de Victoria, A., and Morikis, D. (2011) The effect of electrostatics on Factor H function and related pathologies, *J. Mol. Graph Mod.,* 29(8): 1047-55.
7. Kieslich, C.A., Yang, J., Gunopulos, D., and Morikis, D. (2011) Automated computational protocol for alanine scans and clustering of electrostatic potentials: application to C3d.CR2 association, *Biotech. Prog.*, 27(2): 316-325.
8. Chae, K., Gonong, B.J., Kim, S.C., Kieslich, C.A., Morikis, D., Balasubramanian, S., and Lord, E.M. (2010) A multifaceted study of stigma/style cysteine-rich adhesion (SCA)-like Arabidopsis lipid transfer proteins (LTPs) suggests diversified roles for these LTPs in plant growth and reproduction, *J. Exp. Botany*, 61(15): 4277-4290.