Tracing the Evolution of Electron Transfer (Notes)

Linda Cook

PI: Yana Bromberg

5 May 2015

Abstract

All life needs energy. All life forms fulfill their energy requirements through electron transfer reactions, such as those in cellular respiration and photosynthesis [8]. These electron transfer reactions are mediated by oxioreductase enzymes, whose active components are ancient and of unknown origin. Understanding the evolution of these components will shed light on the history of life.

Since oxioreductases often contain a transition metal as a critical part of their active component [8], we examine the structural similarity of transition metal binding protein folds and determine their evolutionary relationships. Specifically, we obtain all structures from the Protein Data Bank [3] that are involved in binding to a transition metal. We define a 15Å[1] sphere around the transition metals in each protein, and examine only those amino acids within the sphere, the active component. We then perform all-to-all structural alignments of the spheres with transition metals, which we use to define a function to differentiate functionally similar folds from others, based on the number of residues that are aligned and the percent of residues that are identical in an alignment of two proteins. Using this function as edge weights, we constructed a similarity network for all our transition metal spheres. Analysis of this network will help indicate distant relations amongst oxioreductases if any. We further "fished" for parts of proteins in the entire PDB that were structurally similar to our spheres for further analysis.

1 Disclaimer

These are my personal notes and any errors or misrepresentations in them are a result of personal mistakes I may have made, not of the Professor Bromberg's or any of her former or current lab members. If I have incorrectly attributed any work I deeply apologize. I also should that I made use of several scripts Dr. Stefan Senn had previously coded for the lab. Although they would are unlikely to read these notes, I would like to thank Yana Bromberg and Slim Karkar for all their help and advice.

[1] Å means angstrom which is 10^{-10} meters. To put it in perspective a hemoglobin (a type of protein) has about a 25Å radius. Fibrogen, a protein involved in blood clots, is shaped like a 460Å long rod [5]

2 Background

As I am a bioinformatics novice, I tried to write these notes for people like me who don't know a ton of biology and have a background in computing. Before going into the Computer Science components, it is necessary to introduce some biology. Understanding the biological background of this project has been a major part of my learning this semester.

2.1 Oxioreductases

Electron transfer reactions, reactions that involve the movement of one or more electron without the movement of a proton, are critical to life. They fulfill energy requirements of cells and control the biological flux of critical elements to life. On Earth six elements build biological macromolecules, five of which, H, C, N, O and S move biologically because of electron transport reactions [7][6]. These electron transfer reactions are facilitated by oxioreductase enzymes, whose active domains, or components, are believed to have developed more than 2.4 billion years ago and then spread through to all life forms [8]. Thus the history of the domains involved in electron transfer reactions will provide insight into the history electron transfer reactions and life itself.

2.2 Structure vs Sequence

Unfortunately for origins-of-life scientists, the evolutionary history of the domains active in electron transport is very difficult to determine because of their greatly varying amino acid sequences [9]. This project exploits the fact that protein structure is more conserved than amino acid sequence throughout evolution [15]. In fact, a 2009 study has shown that core protein structure is 3-9 times more conserved than sequence, through analysis of sequence identity and a variety of metrics of structural identity of evolutionary related domains [10]. [2] Thus protein structure can be used to determine ancient evolutionary relationships, such as those of the active domains in oxioreductase enzymes [15]. This approach is particularly promising because of the high availability of 3-dimensional protein structural annotations. In the later half of the 20th century, methods to determine the structures were developed [15]. The Protein Data Bank was formed in 1971 to store structural annotations and contained seven structures [2]. Since then, the Protein Data Bank has grown rapidly (Figure 1) and when the data set for this paper was collected there were over 100,000 entries in the Protein Data Bank. Due to the high availability of structural data and its sensitivity to ancient evolutionary relationships, we use structural data to draw conclusions about the development of the active domains in oxioreductases.

Structural comparisons of proteins are limited in that there is no established way to determine the significance of protein relationships [13]. The past work of Dr. Stefan Senn has shown that the distance between transition metal ligands in a structural alignment

[2]There is as yet no consensus of a metric to determine alignment quality [13]. The 2009 paper by Illergård et al., used RMSD and discrete structural descriptors. RMSD, root means squares deviation, is used to find our alignments of protein structures, and is described in greater detail 2.2. In the 2009 paper, core protein sites are considered alignable protein sites, thus it is a relevant result to our study of the alignable active domains in oxioreductases. [10]

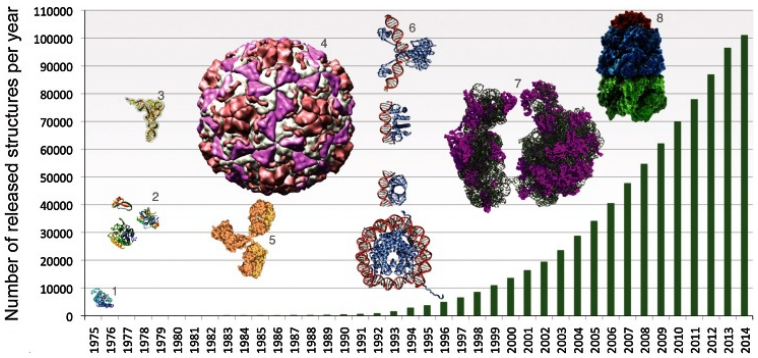

Figure 1: The expansion of the Protein Data Bank with examples of some of the protein structures available in it. Image from [2].

of oxioreducatases can be used to describe the quality of an alignment [13]. A structural alignment is the optimal way to overlay two 3-dimensional protein structures so that they are as close together as possible. This is explained in more detail in Section 2.2.

3 Methods

3.1 Obtaining domains structures of active oxioreductases

3.1.1 Acquiring protein structures containing oxioreductases

From the approximately 100,000 protein structural annotations in the RSCB Protein Data Bank, we took representatives of clusters based on 100% sequence identity. This preliminary step helps to ensure a nonredundant final data set. From these sets we obtained those structures which bind to an atom or a molecule, that is, contained a ligand, and had less than 3Å resolution. The high resolution is important for us to make meaningful comparisons of structures. Then we obtained only those approximately 17,000 structures that contained a transition metal used in oxioreductases, as identified by Dr. Stefan Senn.

3.1.2 Getting the active components

For each of the transition metal ligands in our set, we find every protein in our 17,000 structures set that contains that ligand. We define a 15Å radius sphere from the geometric center of each transition metal ligand in each protein and keep only those amino acids that fall within the sphere (Figure 2). Thus we extract only those folds active in binding to each transition metal ligand, and thus active in electron transport reactions. The extraction was performed by running a script written by Dr. Senn that calculates the Euclidean distance of each amino acid in a protein from the center of the transition metal ligand and writes each amino acid to a sphere file if the distance was less than 15Å. Since this had to be done for each transion metal in each of the 17,000 spheres, we had more than 17,000 independent jobs. The jobs were broken up into batches and run in parallel. A 15Å radius was used because it was the distance found to have the most consistent amount of amino acids across different proteins with different transition metal ligands[13]. As previously shown by John Kim, a 15Å

Figure 2: One of the 4876 metallospheres extracted by drawing a 15 Å radius sphere around the transition metal ligand and keeping only those residues within the sphere. Structure image from PDB[3].

radius also best provides those amino acids involved in binding to the transition metal ligand with the least amount of extraneous amino acids [6]. That is, with too long of a radius the sphere will contain the full protein, including many irrelevant amino acid residues, and with too short of a radius the sphere will not contain many amino acids involved in the binding. We only can make conclusions about spheres that contain sufficient information about the surrounding protein structure, so we removed all those spheres that contained less than 35 amino acids. At this point we also removed the spheres with ligands that contained Zinc, as those were dominated, not by oxioreductases, but by Zinc fingers, another protein structure involved in DNA transcription amongst other things. We ended with a set of approximately 11,500 spheres with transition metal ligands, we called metallospheres.

3.1.3 Removing Redundancy

It is important to note here that some of the 17,000 proteins have multiple transition metal ligands. Since proteins are only around 30 – 50 angstroms [11], in many of theses cases the 15Å spheres overlap. So there were spheres cut from the same protein that contained some of the same amino acid residues (Figure 3). Having spheres with partially identical structures would skew any comparison. So if any two spheres had identical residues, we removed both. These spheres were identified by a python script I wrote over the summer, which simply checked if spheres were from the same protein and, if they were, if they had any amino acids from the same chains and finally if those amino acids were actually the same. The sphere set was reduced to approximately 9,000 in this step.

We now wished to further reduce redundancy for two reasons, the first being the possibility of skewing our data set and the second being reducing the computing power necessary to make all-to-all structural alignments for our spheres. The idea was to put the spheres into groups of high similarity, and keep the member of each group with the highest amount of structural data, that is, the highest number of amino acids. Two spheres were put into the

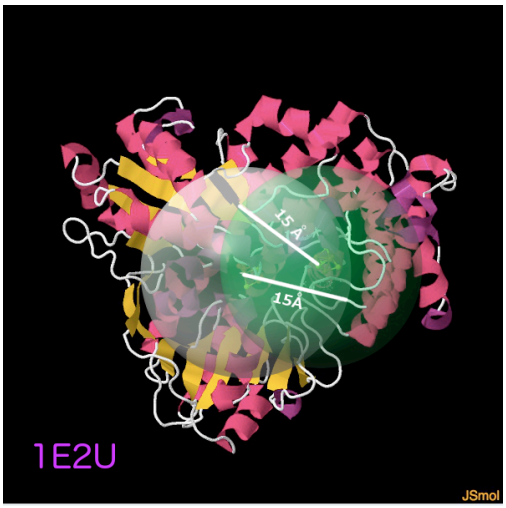

Figure 3: 1e2u, a protein found in a species of sulfate reducing bacteria, was one of our proteins with multiple transition metal ligands. 1e2u's two 15Å radius spheres overlapped, so both were removed from our sphere set. Structure image from PDB[3].

same group if they were centered around the same ligand, had the same number of chains, and those chains fell into the same 100% sequence based clusters (Figure 4). The clustering was performed with BLAST, a tool which follows a much faster, heuristic approximation of Smith-Waterman's sequence alignment algorithm. To divide the spheres into groups my script first put them into metagroups of spheres with the same number of chains and that bound to the same ligand. The cardinality of these metagroups was never larger than 3, so it was possible to check if the chains of members of the metagroups were a part of the same clusters through brute force. Once the spheres were divided into groups we chose the sphere with the greatest number of amino acids and kept it for our final data set of 4,876 metallospheres (Figure 5).

3.2 Comparing Structures

Once we obtained our final set of 4876 active components of oxioreductase enzymes, it was time to begin comparing their structures. We did this by performing all-to-all structural alignments using a tool called Topmatch. Topmatch optimizes the overlay of 3D protein structures based on the root means squares deviation of the distance between the alpha carbon backbones of aligned amino acids and the length of the alignment [14]. Since each alignment is an independent problem, we were able to run them in parallel. For each of the alignments, we took the Euclidean distance between the centers of the two metal ligands in the alignment as a measure of the alignment quality, as per Dr. Senn's work [13] (Figure 6). We defined those alignments with a ligand-ligand distance of 2Å or less as "good" alignments, i.e. alignments implying a high level of biological significance, and the rest as "bad" alignments. We then plotted each alignment according to its length on the x-axis, and the percent sequence identity on the of the aligned portion of the two spheres in the alignment on the y-axis. We optimized the SaHLe (structure-annotated homology, ligand-extended)

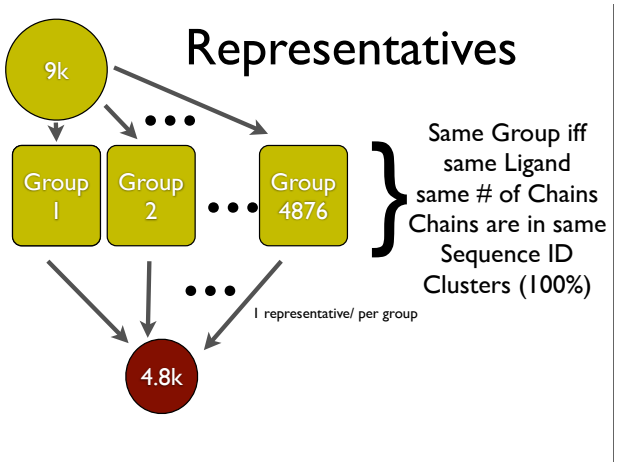

Figure 4: Spheres were put into highly similar groups with the same transition metal ligand, the same number of chains falling into the same sequence based clusters. One representative was chosen per group and added into our final set of 4,876 metallospheres.

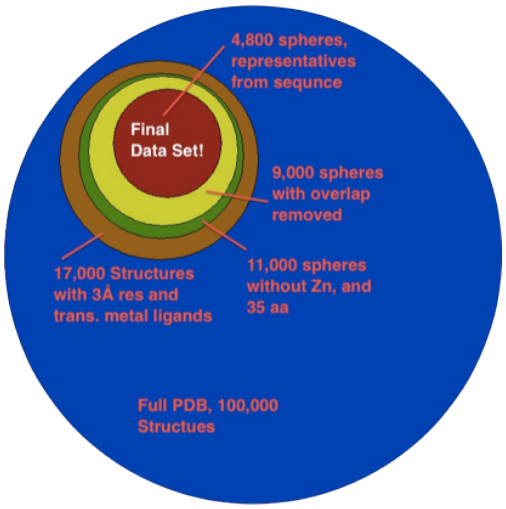

Figure 5: An illustration of the relative cardinalities of the entire PDB (blue) and our intermediate data sets and our final data set (in red) of 4,876 metallospheres.

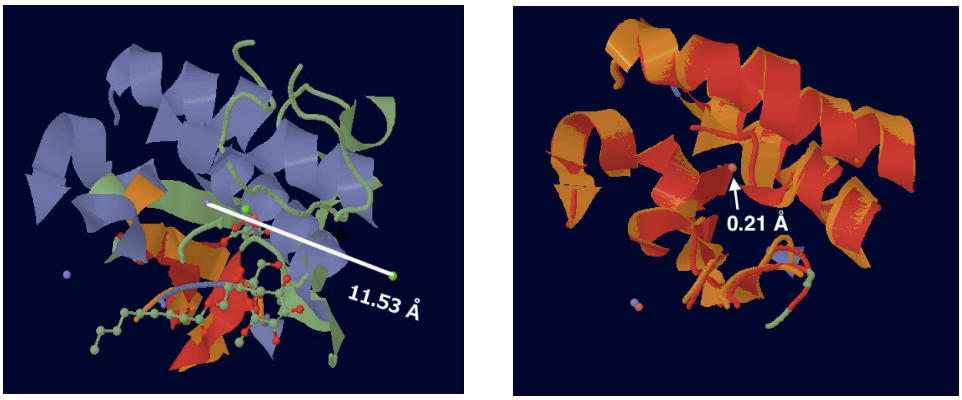

Figure 6: Two alignments of the metallosphere cut from 2cdy are shown. 2cdy's unaligned residues are drawn in blue, and its aligned residues are drawn in orange. The sphere 2cdy is aligned with 3h1s and 1ids have green unaligned residues and red aligned residues. The alignment of 2cdy and 3h1s has a Euclidean distance between metal ligands of 11.53 Å, implying functional disimilarity and that of 2cdy and 1ids has a distance of .21 Å implying the two structures are functionally similar. Structure images from PDB[3].

function to best divide the "good" alignments from the "bad" alignments, based on accuracy and coverage (Figure 7). The resulting formula was $y = b * x^{-a*(1+e^{\frac{-x}{100}})}$ where $a = .598$ and $b = 900$. We defined the SaHLe score, the quality of an alignment, to be the distance from the SaHLe curve along the y (sequence identity) axis to the alignment's point. So the SaHLe score is positive for alignments above the curve, and negative with those below the curve. This fits as the SaHLe curve was optimized to have the good alignments above it and the bad below. Having a Sahle score based on length and sequence identity alone allows us to use it as a metric for spheres that do not contain ligands, such as nonoxioreductase folds we believe may have originated from an oxioreductase ancestor. We kept only those alignments with nonegative SaHLe score, and constructed a similarity network where each vertex was a metallosphere and the edges were weighted by the SaHLe score of the alignment of the two metallospheres it is incident to. Analysis of this structural similarity network should provide insight into evolutionary relationships.

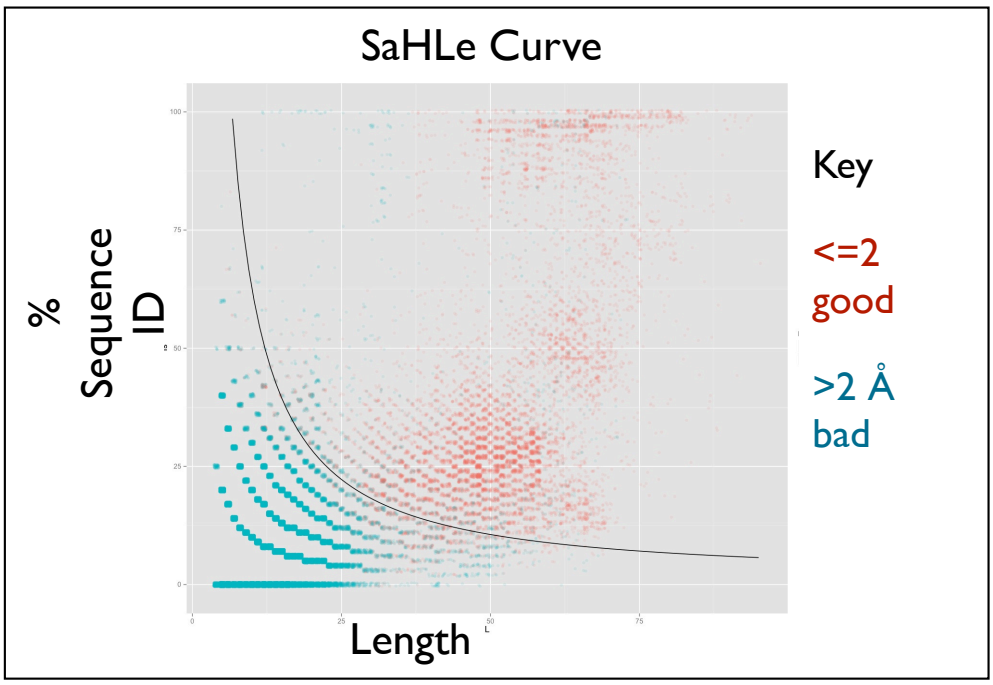

Figure 7: Every alignment was plotted according to its length and the percent sequence identity of the alignment. Then the SaHLe function (drawn in black) was optimized to separate the good (less than equal 2Å ligand-ligand distance) and the other bad alignments.

References

[1] Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11), 3747-3752.

[2] Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. Journal of Computer-Aided Molecular Design, 28(10), 1009–1014. doi:10.1007/s10822-014-9770-y

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

[4] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

[5] Erickson H. P. (2009). Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. Biological procedures online, 11, 32–51.

[6] Falkowski, P. (Director) (2015, March 23). Light to Life. Origins of Life. Lecture presented at the Simons Foundation, New York, New York.

[7] Falkowski PG, Fenchel T, & Delong EF (2008). The microbial engines that drive Earth's biogeochemical cycles. Science 320(5879):1034–1039.

[8] Harel, A., Bromberg, Y., Falkowski, P. G., & Bhattacharya, D. (2014). Evolutionary history of redox metal-binding domains across the tree of life. Proceedings of the National Academy of Sciences of the United States of America, 111(19), 7042–7047. doi:10.1073/pnas.1403676111

[9] Harel, A., Falkowski, P., & Bromberg, Y. (2012). TrAnsFuSE refines the search for protein function: oxidoreductases. Integrative Biology, 4(7), 765-777.

[10] Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics, 77(3), 499-508.

[11] Kuhlman, B. (2008). Amino Acids. Retrieved May 3, 2015.

[12] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

[13] Senn, S., Nanda, V., Falkowski, P., & Bromberg, Y. (2014). Function-based assessment of structural similarity measurements using metal co-factor orientation. Proteins: Structure, Function, and Bioinformatics, 82(4), 648-656.

[14] Sippl, M. J., & Wiederstein, M. (2012). Detection of spatial correlations in protein structures and molecular complexes. Structure, 20(4), 718-728.

[15] Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. Journal of molecular biology, 307(4), 1113-1143.