Mapping Electron Tunneling Pathways: An Algorithm that Finds the "Minimum Length"/Maximum Coupling Pathway between Electron Donors and Acceptors in Proteins

Jonathan N. Betts,[†] David N. Beratan,*[,‡] and José Nelson Onuchic[§]

Contribution from the Beckman Institute, California Institute of Technology, Pasadena, California 91125, Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, and Department of Physics, University of California, San Diego, La Jolla, California 92093. *Received May 13, 1991*

Abstract: The covalent, hydrogen bonded, and van der Waals connectivity of proteins can be represented with geometrical objects called graphs. In these graphs, vertices represent bonds and the connections between them, edges, represent bond–bond interactions. We describe a model in which edge lengths are associated with the wave function decay between interacting pairs of bonds, and a minimum distance graph-search algorithm is used to find the pathways that dominate electron donor–acceptor interactions in these molecules. Predictions of relative electron transfer rates can be made from these pathway lengths. The results are consistent with many experimentally measured electron-transfer rates, although some anomalies exist. Presentation of the pathway coupling between the donor (or acceptor) and every other atom in a given protein as a color-coded map provides a design tool for tailored electron-transfer proteins.

Introduction

Graph theory is often used in chemistry to describe the relationship between molecular structure and chemical properties.[1] Although alternative approaches have been used,[1a,b] traditional chemical graphs consist of a direct mapping of atoms to vertices as well as bonds to edges linking vertices. The graph representations of proteins that are described here employ a slightly different mapping where vertices correspond to bonds and edges to covalent, hydrogen bond, and van der Waals interaction between bonds. The edge lengths represent the wave function decay through these bonded or nonbonded contacts rather than physical lengths. Longer effective lengths represent larger decays. Minimum length pathways often make the dominant contribution to the protein-mediated coupling between electron donor (D) and acceptor (A).

Tunneling Pathways. Many biological reactions shift an electron a considerable distance (>5 Å) via electron tunneling. Such long distance transfers are in the nonadiabatic limit, so the rate is proportional to the square of the protein mediated donor–acceptor coupling, T_{DA}.[2] We recently developed a tunneling pathway model for electron transfer in proteins that identifies the bonded and nonbonded interactions that give rise to the coupling.[3,4] This model is based on an effective one-electron tight-binding hamiltonian. These one-electron interaction parameters are renormalized couplings arising from the more complete hamiltonian. The validity of this reduction and the simple parameter set (discussed below) have been discussed in prior papers.[3d,4d] Using this simple hamiltonian, relative values of T_{DA} have been estimated for a variety of proteins using a single pathway approximation. The single pathway parameters include corrections (in an average sense) due to scattering of electron amplitude in side chains connected to the main pathway.[5a] Alternative electronic structure methods for computing T_{DA} in large systems are being actively pursued.[5] The goal of this method, described in detail here, is to apply the *minimal description* required to incorporate the basic features of the mechanism for electron tunneling in proteins. In spite of these simplifications, this model successfully predicts the relative rates of electron transfer in a large number of experimental systems[4d] and provides the starting point from which the complicating effects of multiple pathways, loop structures in pathways, and many-electron effects can be investigated in a systematic manner.

The algorithm for determining the set of bonds that dominates this D–A interaction is the subject of this paper. Although based on a simple expression for the protein-mediated coupling, the model successfully predicts the relative rates of electron transfer in ruthenated cytochrome *c*,[4,6] myoglobin,[7] and cytochrome b_5.[8] The pathway model explains order of magnitude differences in couplings for specific metal-labeled proteins despite nearly identical D–A separation.[3d,4d,6-8]

The rate of nonadiabatic electron transfer from D to A is

[†] California Institute of Technology. Present Address: Massachusetts Institute of Technology, Mail Stop E34-201, Cambridge, MA 02139.
[‡] University of Pittsburgh.
[§] University of California.

(1) (a) Arteca, G. A.; Mezey, P. G. *Int. J. Quantum Chem.* **1988**, *34*, 517. (b) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. *J. Mol. Biol.* **1989**, *212*, 151. (c) Wilson, R. J.; Watkins, J. J. *Graph Theory, an Introductory Approach*; Wiley: New York, 1990; Chapter 8. (d) Buckley, F.; Harary, F. *Distance in Graphs*; Addison-Wesley: New York, 1990. (e) Maurer, S. B.; Ralston, A. *Discrete Algorithmic Mathematics*; Addison Wesley: New York; 1991; Chapter 3. (f) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic Press: New York, 1976. (g) Trinajstic, N. *Chemical Graph Theory*; CRC Press: New York, 1983. (h) Gutman, I.; Polansky, O. E. *Mathematical Concepts in Organic Chemistry*; Springer-Verlag: Berlin, 1986.

(2) (a) Hopfield, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **1974**, *71*, 3640. (b) Jortner, J. *J. Chem. Phys.* **1976**, *64*, 4860. (c) DeVault, D. *Quantum Mechanical Tunneling in Biological Systems*, 2nd ed.; Cambridge University Press: New York, 1984.

(3) (a) Onuchic, J. N.; Beratan, D. N. *J. Chem. Phys.* **1990**, *92*, 722. (b) Beratan, D. N.; Onuchic, J. N. *Photosynth. Res.* **1989**, *22*, 173. (c) Beratan, D. N.; Onuchic, J. N.; Hopfield, J. J. *J. Chem. Phys.* **1987**, *86*, 4488. (d) Beratan, D. N.; Onuchic, J. N. In *Electron Transfer in Inorganic, Organic, and Biological Systems*; ACS Adv. Chem. Ser. No. 228, Bolton, J. R., Mataga, N., McLendon, G., Eds.; American Chemical Society: Washington, DC, 1991.

(4) (a) Beratan, D. N.; Betts, J. N.; Onuchic, J. N. *Science* **1991**, *252*, 1285. (b) Beratan, D. N.; Onuchic, J. N.; Betts, J. N.; Bowler, B. E.; Gray, H. B. *J. Am. Chem. Soc.* **1990**, *112*, 7915. (c) Beratan, D. N.; Onuchic, J. N.; Gray, H. B. In *Metal Ions in Biological Systems*; Sigel, H., Sigel, A., Eds.; Marcel Dekker Press: New York, 1991; Vol. 27, 97–127. (d) Onuchic, J. N.; Beratan, D. N.; Winkler, J. R.; Gray, H. B. *Annu. Rev. Biophys. Biomol. Struct.* In press.

(5) (a) Onuchic, J. N.; de Andrade, P. C. P.; Beratan, D. N. *J. Chem. Phys.* **1991**, *95*, 1131. (b) Kuki, A., preprint, 1991. (c) Siddarth, P.; Marcus, R. A. *J. Phys. Chem.* **1990**, *94*, 8430. (d) Broo, A.; Larsson, S. *J. Phys. Chem.* **1991**, *95*, 4925. (e) Christensen, H. E. M.; Conrad, L. S.; Mikkelsen, K. V.; Nielsen, M. K.; Ulstrup, J. *Inorg. Chem.* **1990**, *29*, 2808.

(6) (a) Bowler, B. E.; Meade, T. J.; Mayo, S. L.; Richards, J. H.; Gray, H. B. *J. Am. Chem. Soc.* **1989**, *111*, 8757. (b) Therien, M. J.; Selman, M. A.; Gray, H. B.; Chang, I.-J.; Winkler, J. R. *J. Am. Chem. Soc.* **1990**, *112*, 2420. (c) Bowler, B. E.; Raphael, A. L.; Gray, H. B. *Prog. Inorg. Chem.: Bioinorg. Chem.* **1990**, *38*, 259–322. (d) Wuttke, D. S.; Bjerrum, M. J.; Winkler, J. R.; Gray, H. B. *Science*, in press.

(7) Cowan, J. A.; Upmacis, R. K.; Beratan, D. N.; Onuchic, J. N.; Gray, H. B. *Ann. N.Y. Acad. Sci.* **1988**, *550*, 68.

(8) Jacobs, B. A.; Mauk, M. R.; Funk, W. D.; MacGillivray, R. T. A.; Mauk, A. G.; Gray, H. B. *J. Am. Chem. Soc.* **1991**, *113*, 4390.

0002-7863/92/1514-4043$03.00/0

$$k_{ET} = (2\pi/\hbar)|T_{DA}|^2(F.C.) \quad (1)$$

where (F.C.) is the Franck–Condon factor associated with the nuclear motion along the reaction coordinate. A single electron tunneling pathway is defined as a combination of interacting bonds that link D with A via covalent (C), hydrogen bonded (H), or through-space (S) connections. For a single path, the coupling is approximated as[3]

$$T_{DA} \propto \prod_i \epsilon_i^C \prod_j \epsilon_j^S \prod_k \epsilon k^H \quad (2)$$

The goal of the algorithm described here is to choose the combination of bonds between D and A that maximizes the product in eq 2. To provide a simple implementation of the pathway concept, to test its validity, and to show its predictive power, we chose the following parameters:[3,4]

$$\epsilon^C = 0.6 \quad (3a)$$

$$\epsilon^H = 0.36 \exp[-1.7(R - 2.8)] \quad (3b)$$

$$\epsilon^S = 0.6 \exp[-1.7(R - 1.4)] \quad (3c)$$

The distances, R, are in angstroms and the decay factors, ϵ, are unitless. These parameters are consistent with typical binding energies for electron-transfer-localized states as well as theoretical and experimental studies of model compounds.[3] Each decay factor ϵ is associated with an effective distance d_{eff} where

$$d_{eff} = -\log \epsilon \quad (4)$$

We will refer to both decay factors and connection lengths throughout the paper.

Maximum Coupling Pathways. The strength of the coupling arising from a single pathway is proportional to the product of decay factors for each step on the path: $\prod_i \epsilon_i$. The computational challenge before us is to analyze the highly interconnected network of bonded and nonbonded contacts in a protein and specify the bonds that maximize this product. This is precisely the well-known "minimum distance in a graph" problem. The minimum distance problem addresses finding the shortest pathway between two points in an interconnected network. Since eq 4 associates the decay factor with an effective distance, we can restate our search for the maximum pathway coupling as a search for the shortest effective distance between donor and acceptor in the corresponding network. General graph theory strategies for solving the minimum distance problem are discussed in refs 1d and 1e.

Methods

The first step in using graph theory to find electron-transfer pathways in proteins is to construct a labeled graph[1] corresponding to the superset of all interesting potential pathways. Covalent bonds (established as described below) are first mapped onto vertices.[9] Establishing which vertices are to be joined by edges requires progressively more computation for adjacent covalent bonds, hydrogen bonds, and potential through-space (TS) contacts. The lengths of the edges (i.e., the decays) are determined by the distances between the atoms and the nature of the interaction, eq 3. The covalent bonds are specified implicitly by the Brookhaven Protein Data Bank (PDB) files.[10] Covalent interactions, those between bonds anchored at a common atom, are easily identified. Existing software[11] is used to look up these connections for the known amino acids and other residues, which are then appended to the PDB data. Figure 1 outlines the chain of events between PDB file reading and pathway prediction. The degree of connectivity in the resulting graph is shown for a typical protein in Figure 2, and averages about 2.3 connections per atom. These amended PDB files are used as input to the PATHWAYS software[12] written by the authors. On the basis of data in the parameter files, the program looks up the model-predicted decays, eq 3, for the covalent bonds and stores them.

Hydrogen bonds are identified by the following criteria:[13] (1) hydrogen-donor and hydrogen-acceptor groups (donors, $-NH_x$; acceptors, carbonyl oxygens; both, $-OH$); (2) donor–hydrogen–acceptor angle ($\leq 90°$), and (3) donor–acceptor distance (≤ 3.5 Å). These values are specified in a parameter file. Edges representing the hydrogen bonds are

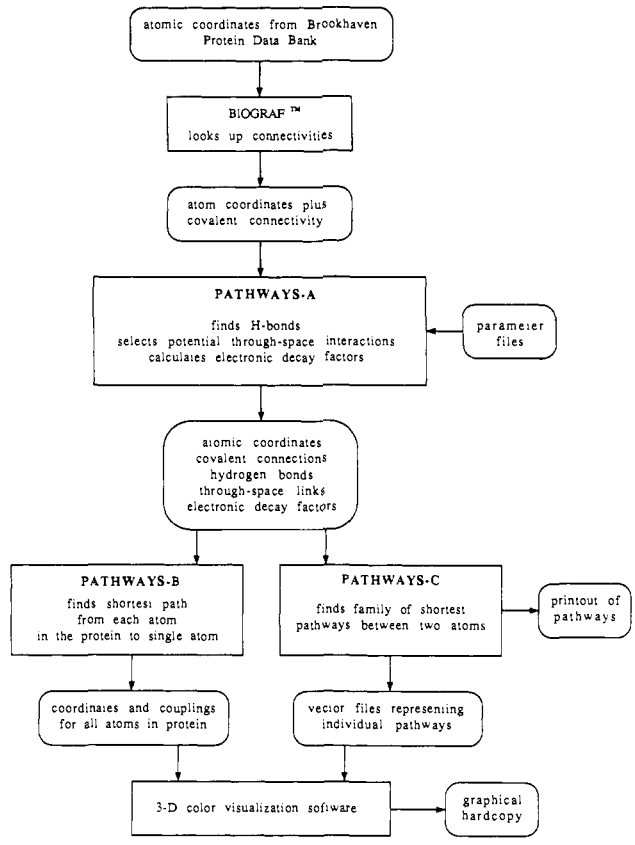


Figure 1. Information flow for calculating electron-transfer pathways in proteins. Rounded cells refer to data and square cells to processes performed by computer programs. PATHWAYS-A, -B, and -C are each part of the PATHWAYS program available from the authors.[12]

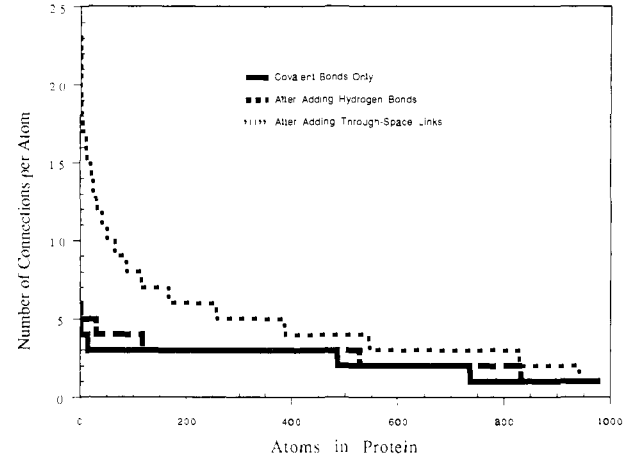


Figure 2. Distribution of connectivity for the heavy atoms in a typical protein (azurin) at various stages in the graph-building process. The connectivities from each stage are sorted on the basis of the number of connections to the atoms.

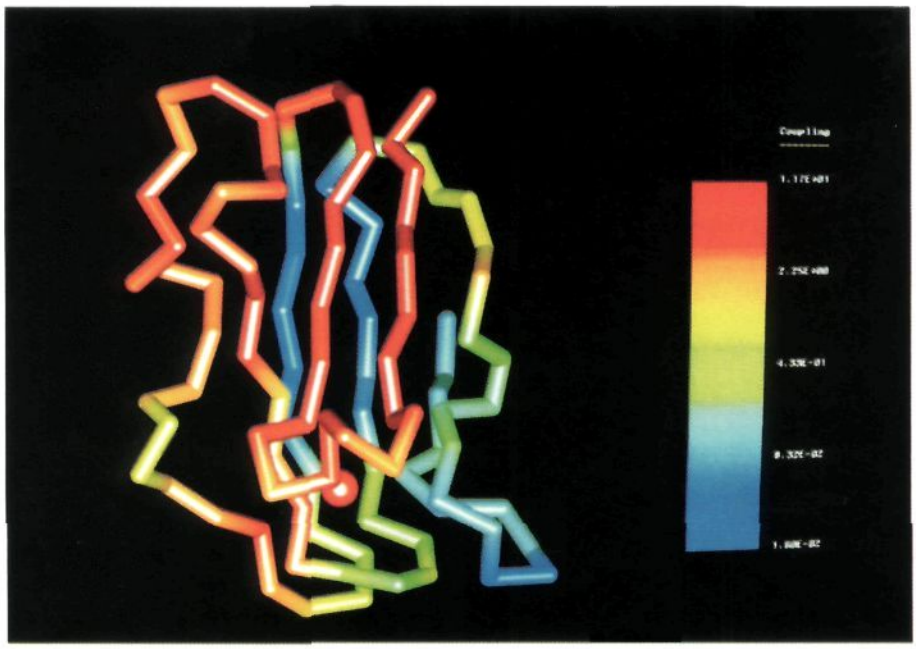
(9) These assignments are an obvious oversimplification. Inaccuracies in treatment of the through-space coupling are introduced by neglecting or adding some lone pair electrons, neglecting hydrogens bound to atoms other than heteroatoms, and suppressing through-space orientation effects. However, if all of these effects were included, corrections to the overall decay would likely be of order unity because there are very few through-space connections in the dominant paths. Errors in the through-space decays do not affect any of the qualitative predictions of the pathway analysis.

(10) Bernstein, F. C.; Koetzle, T. E.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, M.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.

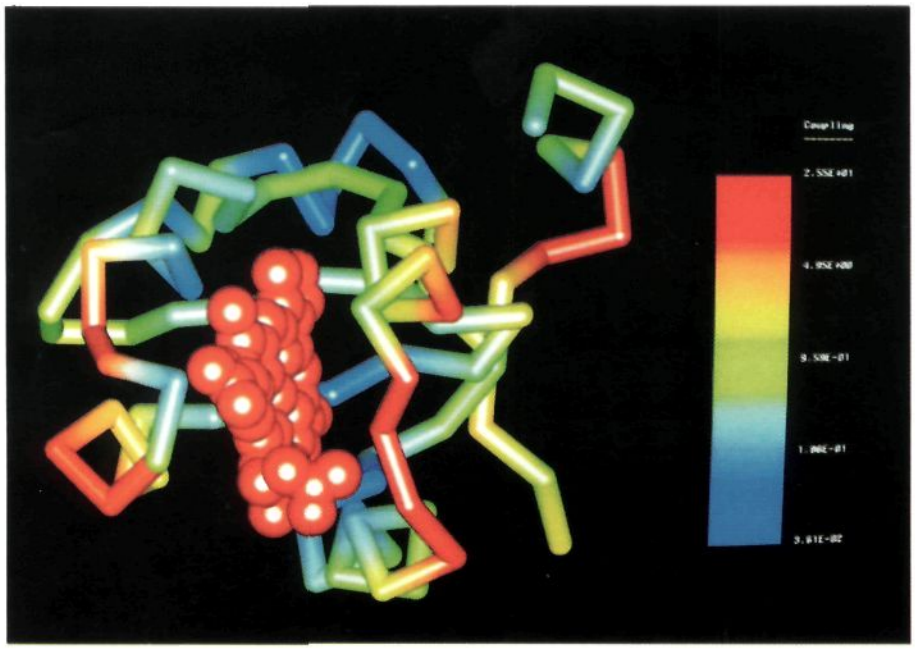
(11) For example, BIOGRAF, a product of Biodesign, Inc., Pasadena, CA 91101, was used here.

(12) The software (PATHWAYS v. 2.2) and user's manual are available from D.N.B.

(13) See, for example: Stryer, L. *Biochemistry*, 2nd ed.; W. H. Freeman and Co.: New York, 1981.



(a) Plastocyanin



(b) Cytochrome b5

Figure 3. Pathway coupling ratio maps are shown for (a) plastocyanin and (b) cytochrome b_5. Note that the antiparallel β-sheet (barrel) structure in plastocyanin provides "hot" spots in the strands ligating the Cu center but not in the other strands. In cytochrome b_5, however, the β-sheet structure (shown here behind the heme in a plane roughly perpendicular to it) does not radiate from the porphyrin, so it does not assist coupling along the full length of the protein as it does in plastocyanin. Displayed is $\prod \epsilon_i/[A \exp(-R\beta/2)]$ where the numerator is the pathway mediated coupling to an α-carbon and the denominator is the best fit exponential expression for T_{DA} for all α-carbons in the entire protein, evaluated for each α-carbon at distance R from the heme or Cu site.

added to the connection list, and the lengths that represent these decays are added to the list of segment lengths. This increase the degree of connectivity for the protein by about 0.25 per atom (see Figure 2).

Potential through-space (TS) connections are sought within a limited radius of each atom, typically 6 Å. It was found that no TS connections longer than this contributed to significant pathways, so the irrelevant long distance ones beyond this cutoff distance are eliminated to shorten the data-processing time. The TS connections are established for each atom, **A**, as follows. First, a list, **L**, containing all bonds/vertices within range of **A** is made and an attempt is made to eliminate as many of the entries as possible. Eliminating the TS connections between two atoms having a significantly better alternative through-bond connection was found to decrease the average added connectivity from about 21.3 to 1.9 per atom. The first connections the program eliminates from **L** are those that are redundant with preexisting covalent and hydrogen bonds. The vertices remaining in **L** are sorted on the basis of their distances from **A**, shortest first. Next, a depth-first shortest-path search[1] is performed (see next section) with **A** as the root, finding the shortest distance to **F** (the first

vertex in **L**) through the already existing connections. The depth of the search is limited to a length which corresponds to the TS decay from **A** to **F**. If the search returns without having located **F**, then the TS contact is the shortest path, and is thus added to the master connection (adjacency) list, and its corresponding length is added to the list of lengths, otherwise, **F** is discarded. Then the next vertex in **L** becomes the new **F**. In this way, shorter TS contacts may contribute to favorable paths and can disqualify longer ones, further decreasing the amount of connectivity added to the graph. The resulting change in connectivity is shown in Figure 2.

The Search Algorithm. There are two standard search strategies for arriving at the minimum-distance path between two points in an interconnected network, referred to as depth-first and breadth-first searches.[1] A depth-first search begins at a specified point and steps along allowed connections until no additional forward steps exist (a dead-end is reached) or the target site is found. If a dead-end occurs, the search backtracks by one step and then seeks alternative forward steps from that point, and so on until the target atom is found. A breadth-first search simultaneously considers all paths radiating from the starting point by keeping track of each vertex and its distance. At each step of the search a new vertex is added. The vertex chosen to be added is always the one that minimizes the effective distance to the donor at that stage. When the acceptor atom is the one that is added, the minimum distance pathway has been found. We use a depth-first algorithm in this work. The advantage of the depth-first search for our application is its "pathway orientation", i.e., each excursion represents a potentially acceptable pathway and the paths within a given factor of the best one are easily tabulated and accumulated.

Once the adjacency list is complete, and the lengths of the edges are calculated, the graph is ready for shortest-path searches to be executed. The depth-first search algorithm used in PATHWAYS can be described recursively in approximate terms as follows:

```
begin SEARCH(base, length)
  ispath(base) = true
  branch = 1
  probe = adj(base, branch)
  do while probe ≠ 0
    if length + len(base, probe) < sofar(probe) and ispath(probe) =
          false then
      call SEARCH(probe, length + len(base,probe))
    else
      branch = branch + 1
      probe = adj(base, branch)
    end if
  end do
  ispath(base) = false
return
```

where len(i, j) is a 2-D array containing the length of the jth edge connected to vertex i, sofar(i) is an array with the length of the shortest approach made to vertex i so far, ispath(i) is an array that notes whether or not vertex i is part of the currently searched path, and adj(i,j) is the adjacency list, a 2-D array holding the number of the jth vertex connected to vertex i.

After SEARCH(root, 0) is called, the sofar(i) array contains the shortest pathway from root to all other vertices (see Figure 1). In our actual implementation, recursion is not used, and a stack is explicitly maintained. This allows the pathways to be recorded mid-search, and the search to be terminated more easily.

SEARCH is used several times in the program. During the process of locating and eliminating TS connections, the search routine sets a flag and returns immediately if a path to a given target atom is found. The sofar array is not erased between searches from a given atom, so the searches accelerate progressively.

Searches are executed between two atoms or from one atom to all others in the protein. During searches of the entire protein, the routine is allowed to run to completion. The sofar(i) array is used to generate statistics such as a regression of the pathway-based couplings versus through-space distance. The sofar(i) array is also incorporated in output files which are used by our custom graphics-display software to view the electronic couplings as color-coded maps (Figure 3).

For searches between specific atoms, the routine is allowed to run to completion, *but* is interrupted whenever the target atom is encountered in order to record the current pathway. The criteria in this search are relaxed using a sloppiness parameter so that all paths within a variable factor of the best one are retained. Branches are only skipped if the length accumulated to reach them is longer than that atom's entry in the sofar(i) array *minus* the length specified by the sloppiness parameter. In this way, nearly equivalent pathways will not prevent one another from being found. Thus, families of pathways are recorded. After the call to SEARCH, pathways and their lengths are output as tabular reports and as graphics-compatible files.

Discussion

We have described a search algorithm to find electron tunneling pathways with maximal coupling given a simple prescription for through-bond and through-space electronic decay. The method has been used with success to predict relative rates of transfer in several transition metal labeled proteins. The capability of performing global searches for best pathways in a protein from a single site (for example donor or acceptor) to all heavy atoms allows (1) the construction of global protein coupling maps, (2) the identification of "hot" and "cold" spots[4a] for electron transfer at a given distance, and (3) determination of secondary and tertiary motif effects on the coupling. [Hot and cold spots are defined by fitting the pathways couplings for every site in a specific protein to a single exponential in distance to determine the average decay. Sites that are coupled more strongly (weakly) compared to the average value for that distance are termed hot (cold).] Equipped with improved bond and orientation dependent ϵ values, the algorithm could provide lists of the lowest-order perturbation theory pathways for a given level of electronic structure theory (e.g., extended-Hückel).

A key test of the theory involves the attachment of transition metal probes to residues at similar distances that are predicted by the pathway model to have vastly different coupling.[6] The blue copper proteins are systems in which dramatic effects are predicted. Figure 3 shows hot and cold spots in plastocyanin and cytochrome b_5. In plastocyanin, hot spots radiate from the β-strands ligating the Cu. Shorter-range hot spots in cytochrome b_5 are associated with amino acid/heme hydrogen bonding interactions. By changing the protein interactions with the redox centers or by modifying the electronic structure of the ground/excited states (heme or Cu orbitals), it should be possible to change the rates as well (via the prefactors not explicitly in eq 2).

The pathway method has pointed to anomalous electron transfer rates in some systems,[4a,6c] which are now being investigated in further detail experimentally. Further application of this search algorithm should provide a deeper understanding of electron transfer reactions in proteins and nucleic acids, and the manipulation of pathways may also allow the design of stabilized high-energy charge-separated species for more efficient energy conversion schemes.[14] The method is now being refined to include multiple interfering pathways and bond type differences.[5a]

Programming Environment

The software [12] was developed using Silicon Graphics FORTRAN under the IRIX (UNIX) operating system on a Silicon Graphics IRIS 4D/210 VGX. The software will run on any Silicon Graphics IRIS, and should be portable to most UNIX systems supporting FORTRAN. BIOGRAF [11] is used to create the covalent list, but this file could also be generated by other means. Timing for albacore cytochrome c on the 4D/210: 50 sec to construct connection list; 9 sec to calculate all best paths to the heme.

Acknowledgment. This work was performed in part at the Jet Propulsion Laboratory, California Institute of Technology and was sponsored by the Department of Energy's Catalysis/Biocatalysis Program (Advanced Industrial Concepts Division), through an agreement with the National Aeronautics and Space Administration. J.N.O. thanks the National Science Foundation (Grant No. DMB-9018768) and the Department of Energy's Catalysis/Biocatalysis Program (through a research contract from the Jet Propulsion Laboratory) for support of this work. The pathway search software, written in the FORTRAN for Silicon Graphics IRIS computers, is available from D.N.B. J.N.O is in residence at the Instituto de Física e Química de São Carlos, Universidade de São Paulo, 13560, São Carlos, SP, Brazil, during the summers.

Registry No. Cytochrome b_5, 9035-39-6.

(14) Beratan, D. N.; Betts, J. N.; Onuchic, J. N. *J. Phys. Chem.* In press.