

# Evolutionary Trace analysis of dihydrofolate reductase (DHFR)

---

Dr. Teresa Lyons, Accelrys, demonstrates here that Evolutionary Trace is an accurate and thorough application to highlight functionally important residues in dihydrofolate reductase (DHFR).

## Introduction

DHFR reduces dihydrofolic acid to tetrahydrofolic acid. Folate shuttles carbon atoms to various enzymes that need them in their reactions. In one example, DHFR is essential in the pathway for DNA synthesis. While the DHFR structure is generally conserved throughout nature, the mammalian form is about 3000 Daltons larger than the bacterial form, most of the additional residues occurring in loops. The amino acid sequences are quite diverse across species, being as little as 30% identical between bacterial species. Eukaryotes and prokaryotes only share about 30% identity at best. These distinctions have made DHFR an excellent candidate for species-specific drugs such as the antibiotic trimethoprim and the antimalarial pyrimethamine. DHFR is also the target of the chemotherapeutic methotrexate as well as the first anti-cancer drug aminopterin.

Evolutionary Trace, developed by Olivier Lichtarge from Baylor College of Medicine (Lichtarge et al. (1996), *J. Mol. Biol.*, **257**, 342-358), exploits the fact that residues important to the structure or function of a protein strongly tend to be conserved across species. The method uses a multiple sequence alignment of a protein family to identify conserved residues within the family. The Evolutionary Trace algorithm analyzes the conservation of amino acids within a protein family and maps this information in the context of an atomic structure. Hence, spatial clusters of evolutionarily important residues can be identified.

This project is possible because the amino acid sequences of DHFRs from many species are known and there are a number of representative crystal structures of the protein. The protein has a very large groove where it co-localizes NADPH (the hydrogen donor for the reaction) with dihydrofolate (DHF). The roles of several of the residues in stabilizing the hydrogen transfer can be inferred by looking at NADPH and folate in a crystal structure (PDB ID 1RA2).

Some general interactions can also be inferred from a movie of the enzyme that is a composite of six crystal structures of *E. coli* DHFR with compounds bound that mimic various stages of the kinetic pathway (<http://chem-faculty.ucsd.edu/kraut/dhfr.html>). An Evolutionary Trace analysis of DHFR accurately identifies the residues lining the binding pocket and picks out specifically the amino acids that appear to play an important role in binding NADPH and DHF.

## Methods

Discovery Studio (DS) Modeling 1.1 was used for all aspects of this project. Other DS applications mentioned below are part of the DS Modeling suite of programs.

### Preparation of the protein

Several structures of *E. coli* DHFR with different inhibitors with and without NADPH have been elucidated by x-ray crystallography. For this project, I chose 1RA2, a high resolution (1.6 Angstrom) structure that has the whole cofactor (NADPH) and a whole inhibitor (folate) modeled into the active site. I fixed the bond orders on both small molecules and added hydrogens, adjusting the charge on the folate to be negative two as the PDB header indicates it should be.

To optimize the hydrogen positions, I used DS CHARMM to energy minimize the hydrogens, holding the rest of the complex with harmonic constraints. I followed this with a similar minimization, this time allowing the NADPH and folate to be unconstrained. This resulted in slight (less than 1 Angstrom RMSD) movement of the ligand and cofactor, optimizing their contacts with each other and the protein.

### Preparation of the sequences

I used *E. coli* DHFR as the query sequence for a gapped BLAST search of the nr (non-redundant) sequence database at NCBI, initiating the job using DS Similarity Search. Sequences in the BLAST hit list that appeared unique and wild type based on the sequence description (different organisms and unmutated) were selected down to 35% identity and loaded into the sequence window. These sequences were renamed based on the organism the sequence came from (again, as read from the sequence description). The set of sequences were aligned using Align123. The alignment was then adjusted by hand as needed in the sequence window, which involved moving some gaps a space or two in either direction. This alignment was then used to create a sequence dendrogram and duplicate sequences (>98% identical to another sequence) were removed. One sequence, from *S. flexneri*, was removed because it was 20 residues shorter in the N-terminus than the rest of the sequences and prevented some highly conserved residues from being picked up by the trace.

### Evolutionary Trace

The resulting sequence set contains 62 dihydrofolate reductase sequences. The curated alignment described above was used to generate an Evolutionary Trace sequence dendogram and residue clusters using DS Protein Families. A sliding vertical line on the sequence dendogram interactively highlights the trace residues in the structure and sequence windows. The residue clusters were created at a 70 percent identity cutoff (PIC) using the *E. coli* DHFR (PDBID 1RA2) as the clustering structure.

### Determination of important residues in DHFR by visual inspection

The 1RA2 PDB structure was analyzed through the molecule visualization window of DS Modeling. The DHFR-NADPH-folate complex shows several intermolecular hydrogen bonds and hydrophobic interactions.

### Folate interactions with the protein

The folate forms five hydrogen bonds to the protein: two of these are to Asp27, two to Arg57, and one to Lys32. Folate also hydrogen bonds to a buried water molecule that forms a hydrogen bond bridge to Trp30 and Thr113. The hydrophobic interactions include an aromatic interaction to Phe31. Several hydrophobic aliphatic residues also line the binding pocket: Leu28, Ile50, and Leu54.

### NADPH-DHFR interactions

The NADPH molecule forms nine hydrogen bonds to the protein. Seven of these are sidechain hydrogen bonds: two to Arg44, two to Gln102, and one to each of Thr46, Ser63 and Ser64. The remaining two are to backbone atoms for residues Ile14 and Ala7. No specific hydrophobic interactions seem to exist (which is logical since this is a highly polar molecule). However, when the protein is rendered in CPK (space-filling model), several residues occlude NADPH from leaving the binding pocket: Glu17, Asn18, Ala19 and Asp122.

### Crystallographic movie

An internet search led me to a web site entitled "DHFR the Movie" (<http://chem-faculty.ucsd.edu/kraut/dhfr.html>). This animation of how NADPH and DHF exchange into and out of the large binding pocket of DHFR is based upon six crystal structures of DHFR with different transition state analogs bound. Two parts of the molecule act to open and close the binding pocket: the M20 loop at the bottom of the pocket, consisting of residues M16 through L24, and the helix above the pocket, H45-I50. Trp22 and Pro21 anchor the M20 loop while other, more polar, residues swing in and out of the pocket as NADPH and DHF bind and are released. On the helix, His45, Thr46, Ser48 and Ser49 all point somewhat into the pocket.

## Results

### Identification of trace residues

Typically, conserved residues over a sequence family of this size are residues that either form the hydrophobic core of the protein or are absolutely critical to activity of the protein family. Glycines, which are important for transitions between elements of secondary structure and flexibility of loops, also tend to be well represented in the conserved residues, as they are here.

As can be seen from the table below, four of the twelve conserved residues are glycines. All the other residues except M42, which is buried in the hydrophobic core, play an important role in the binding of the two compounds, either through direct contacts or flexibility of the pocket.

**Table 1. Conserved residues in the sequence alignment (*E. coli* DHFR as reference)**

	Contacts Folate	Contacts NADP	M20-loop (16-24)	Helix (45-50)	<10% Solvent-Exposed
A7	.	X	.	.	.
I14	.	X	.	.	.
G15	.	.	.	.	.
W22	.	.	X	.	.
F31	X	.	.	.	.
M42	.	.	.	.	X
G43	.	.	.	.	X
R44	.	X	.	X	.
L54	X	.	.	.	.
R57	X	.	.	.	.
G95	.	.	.	.	X
G96	.	.	.	.	.

Class specific residues are residues conserved within a group of sequences in the family at a particular percent identity cutoff (PIC). These tend to be involved in modulation of function within a family of sequences. Often these residue substitutions are conservative, resulting simply in a slightly longer or shorter hydrogen bond being formed or a slightly different shaped hydrophobic interaction surface. These kinds of changes can be rationalized to result in the observed differences in activities between the family members. They can also be exploited by targeted drug design when differences occur clearly across phylogenetic lines.

The table below lists the residues that are class specific in this family of sequences at various PICs. The most critical residues in activity should be most conserved (lower PIC).

**Table 2: Class-Specific Trace residues at increasing PICs and their inferred roles from the *E. coli* DHFR structure**

	Percent Identity Cutoff	Contacts Folate	Contacts NADP	M20-loop (16-24)	Helix (45-50)	<10% Solvent-Exposed	Alternate residues
T35	< 40					X	S
T46	< 40		X		X		V
T113	< 40	X				X	V I
F125 <sup>1</sup>	40 - 50					X	L M V A
P21	50 - 60			X			A
D27	50 - 60	X					E
K38 <sup>2</sup>	50 - 60						H A R Q G E
S49	50 - 60				X		T
N59	50 - 60					X	S H
Y100	50 - 60					X	F I
A6 <sup>3</sup>	60 - 70					X	
N18	60 - 70		X	X			G Q T A
M20	60 - 70			X			
L24	60 - 70			X			V I Q
L28	60 - 70	X					F I M Q
W30	60 - 70	X				X	H L R Y F
K32	60 - 70	X					
P39	60 - 70					X	
I50	60 - 70	X			X		L
G51	60 - 70						
P53	60 - 70						L V A
S63	60 - 70		X				
R71	60 - 70						
A81	60 - 70					X	V S L
V93	60 - 70					X	
G97	60 - 70					X	
Y111	60 - 70					X	
I115	60 - 70					X	
A117	60 - 70					X	
G121	60 - 70						D A T L
D122	60 - 70		X				N P A
T123	60 - 70						
Y151	60 - 70					X	

1. Pink highlighted rows are residues that are buried and do not have any other inferred structural role.
2. Yellow highlighted rows are trace residues which do not fit into any role categories.
3. Cyan highlighted rows are trace residues in the 60-70% group that have a PIC of higher than 66.26. Most of these residues have high variability and their alternate residues are not listed.

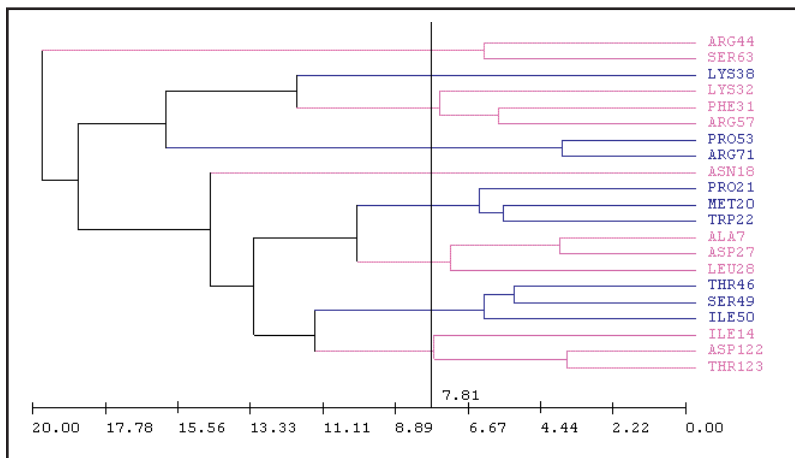
To summarize Tables 1 and 2, almost a third of the amino acids of E. coli DHFR (45 of 159) come up as trace residues at 70 or less PIC. Of the 45 trace residues, 17 are less than 10% exposed to solvent and probably play a role in stabilizing the protein core.

Every residue that interacts directly or indirectly with the folate in 1RA2 is represented in the trace residues. Three of the seven are absolutely conserved over the sequence family. The majority of the seven residues that hydrogen bond to the NADPH molecule show up as trace residues: Ala7, Ile14, Arg44, Thr46, and Ser63. In addition, two of the four residues that occlude the pocket when the protein is CPK rendered are part of the trace: Asn18 and Asp122. Three of the residues of the mobile helix at the top of the binding pocket (residues 45-50) and five of the nine residues in the M20 loop are highlighted in this analysis as well.

Only 7 of the 45 trace residues play no apparent role in DHFR function; four of these are glycines and can be considered important for the protein's general flexibility. At a PIC of 60 and less, only one amino acid (K38) does not play a clear role in the binding of NADPH or the DHFR ligand simply by observation of the protein structure. This residue also has relatively high variability within the family, being preserved essentially as a small aliphatic or charged sidechain, which is consistent with its position on the surface of the protein.

### **Spatial clustering of the residues**

Up to this point, the analysis of DHFR simply demonstrates that the majority of the residues that interact with the cofactor (NADPH) and a ligand analog (folate) show up as being important by simple sequence analysis. One could infer that the sequence family alone is sufficient to prioritize amino acids to target for functional modification by mutagenesis. However, Evolutionary Trace adds value to the simple sequence alignment by including a representative structure from the family and allowing the user to cluster the critical residues spatially. Patches of surface exposed conserved residues tend to play common roles within a protein class. Evolutionary Trace produces a residue cluster dendrogram where the residues are the y-axis and the distance separation is the x-axis. In Discovery Studio Modeling, the residues clustering is not limited to just the trace residues; it may be further culled using other user-defined subsets. For Figure 1, glycines and residues that are less than 10% solvent exposed have been removed from the clustering. Spatial clustering of the trace residues without filtering out the buried amino acids yields several larger clusters, most of which are essentially in the core of the protein.



▲ Figure 1. Spatial clustering dendrogram of the trace residues at 70 PIC filtered to include only solvent exposed (greater than 10% solvent accessible), non-glycine residues. The x-axis is the pairwise distance between residue sidechain centers.

The value of spatial clustering of trace residues on DHFR is immediately obvious: the clusters that line the pocket delineate nicely where the enzyme reaction occurs. If we had only a crystal structure of DHFR with NADPH bound, this analysis would clearly tell us that something else should be binding at the far end of the pocket. Conversely, if no NADPH were present, we'd need to ask the question of what important role is played by the yellow cluster residues. Additionally and perhaps of equal importance, it is clear that there are no other sites on this protein that are involved in its activity: no clusters of residues appear on any other side of the protein.

In conclusion, the Evolutionary Trace method clearly identifies clusters of functionally important residues, outlining the large and complex binding pocket of DHFR thoroughly while making it clear that there are no other areas on DHFR that are essential for its function.



▲ Figure 2. Mapping of the surface-exposed, non-glycine trace residues on DHFR. These residue clusters are involved predominantly in making contact below and to the side of folate (red and green), bridging between the cofactor and ligand at the top and bottom of the pocket (orange and blue) and making contact with the NADPH (yellow). The folate is shown in stick rendering; the NADPH is shown in ball-and-stick rendering.

Five small (3-residue) clusters are revealed at about 7.8 Angstrom separation (Fig. 2):

Cluster 1 (red): K32, F31, R57

Cluster 2 (blue): M20, P21, T22

Cluster 3 (green): A7, D27, L28

Cluster 4 (orange): T46, S49, I50

Cluster 5 (yellow): I14, D122, T123