Published in final edited form as:

Biochemistry. 2010 May 18; 49(19): 4003–4005. doi:10.1021/bi100445g.

Studying enzyme-substrate specificity in silico: A case study of the *E. coli* glycolysis pathway

Chakrapani Kalyanaraman and Matthew P. Jacobson *

Department of Pharmaceutical Chemistry, University of California, San Francisco, 94158-2517

Abstract

In silico protein-ligand docking methods have proved useful in drug design and have also shown promise for predicting the substrates of enzymes, an important goal given the number of enzymes with uncertain function. Further testing of this latter approach is critical because 1) metabolites are on average much more polar than drug-like compounds, and 2) binding is necessary but not sufficient for catalysis. Here, we demonstrate that docking against the enzymes that participate in the 10 major steps of the glycolysis pathway in *E. coli* succeeds in identifying the substrates among the top 1% of a virtual metabolite library.

Assigning protein function based on sequence or structure is remarkably difficult (1,2). Even if two proteins are highly homologous to one another and have similar structures, a change of only a few residues in the active site can change the functional specificity (1,2). We and others have taken a computer-aided, structure-based approach to investigate the *in vitro* substrate specificity of enzymes (3–12). In brief, we have used computational docking methods in conjunction with enzyme structures or homology models to suggest possible substrates for experimental testing. This work is predicated on a hypothesis that specificity of enzymes for their substrates is achieved, in part, through binding specificity, to the extent that most small metabolites the enzyme encounters do not bind in the active site with significant affinity. Substrate binding is clearly necessary but not sufficient for a metabolite to be a substrate.

Our experience with applying the computational metabolite docking approach in both retrospective (3) and prospective (10–12) tests to the alpha-beta barrel enzymes in the enolase superfamily has suggested that this approach is viable and useful in practice. This experience has paralleled that of other groups who have focused on other systems, using similar but distinct computational methods. For example, Shoichet and co-workers have reported successful retrospective (5) and prospective (6,7) tests in the amidohydrolase superfamily, another group of alpha-beta barrels. However, overall there has been far less testing of docking and scoring methods for enzyme-substrate recognition than there has been for the binding of drug-like molecules to a great many drug targets. Further testing of this approach is particularly important because success can be limited not only by the usual challenges associated with sampling and scoring, but also by the underlying assumption that predicted substrate (or enzymatic intermediate (5), in cases where that approach can be applied) binding can be employed as a useful filter to suggest possible substrates.

600 16th St., Box 2240, San Francisco, CA 94158-2517, matt.jacobson@ucsf.edu, phone: (415) 514-9811, FAX: (415) 502-4222.

SUPPORTING INFORMATION AVAILABLE

Detailed information about test set and computational methods, and predicted docking poses. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Here, we use the glycolysis pathway as a case study for investigating whether computational methods can profitably identify potential substrates. We judge success by two criteria: 1) the ability to rank the known substrates among the best scoring metabolites out of a large library, and 2) the ability to distinguish the correct substrate for a given enzyme among other metabolites in the same pathway (and vice versa, i.e., identify the correct enzyme for a particular metabolite in the pathway). The latter is a challenging test of the ability to capture specificity, because the various chemical species in a pathway are of course closely chemically related. These results thus complement our previous work where we evaluated the ability to identify the correct substrate-enzyme pairs among the enzymes within the functionally diverse enolase superfamily (3,10–12). In that case, the substrates were chemically diverse but the enzymes were very similar, at least at the backbone level (13). Here, the substrates are chemically relatively similar but the enzymes represent many different superfamilies. Specifically, the pathway includes 4 kinases, 2 isomerases, a dehydrogenase, an aldolase, a mutase, and an enolase.

The computational methods have been described in detail previously (3). Briefly, we used Glide (14) to dock a library of ~19k metabolites and other biologically active compounds taken from KEGG (15) against structures or homology models of the 10 enzymes listed in Table 1. (See Supplementary Information for detailed methods.) With the exception of phosphoglucose isomerase (step II), the lowest-energy ligand-binding pose predicted by Glide closely mimicked that in the crystal structure of the enzyme or the template structure used for the homology model (Figure S1). The phosphoglucose isomerase structure (pdb id 2crr) was co-crystallized with a linear form of 6-phosphogluconic acid. Although the metabolite library contained both linear and cyclic forms, the cyclic form received a better score. Interestingly, however, phosphoglucose isomerase is believed to catalyze the ring opening of the cyclic substrate (16).

Ranks of substrates, expressed as percent of the total metabolite library, are given in Table 2 for all 10 enzymes in the glycolysis pathway. Enzymes that participate in each step of the glycolysis are presented in the rows, and the substrates of each enzyme are given in the columns, so that the diagonal matrix elements represent cognate enzyme-substrate pairs. The docking program identified the cognate substrates within the top 1% for 3 enzymes: triosephosphate isomerase (V), phosphoglycerate kinase (VII) and phosphoglycerate mutase (VIII). Cognate substrates of 5 other enzymes were identified in the top 5% and the remaining ones were found within the top 13%.

In previous work, we have found that applying a post-docking rescoring procedure can significantly improve the results of metabolite docking (3). Specifically, we use a molecular mechanics force field in combination with a generalized Born implicit solvent model (MM-GBSA) to energy minimize the ligand in the binding site and rank the ligands according to predicted relative binding affinity.

The results in Table 3 show that the MM-GBSA rescoring procedure significantly improves the ranks of cognate substrates, which are all within the top 1%. In 6 out of 10 cases, cognate substrates are ranked within the top 0.3%, i.e., among the top scoring ~50 ligands. In addition, when the docking program has already found the cognate substrate in the top 1% cutoff, performing MM-GBSA rescoring improved the ranking further.

Because the product of one enzyme is the substrate of the next enzyme in the pathway, the underlined, immediately off-diagonal matrix elements represent the ranks of reaction products. In most cases, the products rank highly, in some cases even slightly outranking the substrate. There are two dramatic exceptions: glucokinase (I) and glyceraldehyde 3-phosphate dehydrogenase (VI). In both cases, we used a structure (or model based on a structure) co-

crystallized with reactant. We suspect, but cannot prove, that conformational changes in the binding site would be necessary for product to rank highly. In the case of glucokinase (I), the predicted binding pose of the product glucose 6-phosphate places the glucose moiety in a position similar to the substrate causing the phosphate group to have unfavorable electrostatic repulsion with Glu187. In the case of glyceraldehyde 3-phosphate dehydrogenase (VI), the predicted docking pose of the product is incorrectly 'flipped' whereas the substrate docked correctly. It should be noted that, while the MM-GBSA scoring function performs well in ranking substrates highly, it is more sensitive to such atomic level detail than the more empirical scoring function used in the docking program.

The other off-diagonal matrix elements provide information about the ability to discriminate the correct substrate for the correct enzyme within the pathway, a challenging test of the ability to capture selectivity. Examining each row in Table 3, we note that for 3 enzymes—glucokinase (I), phosphofructokinase (III) and enolase (IX)—the cognate substrates rank ahead of all other glycolysis pathway metabolites. In 2 other cases, triosephosphate isomerase (V) and phosphoglycerate mutase (VIII), the product of the reaction ranks the highest. In the other cases, the cognate substrate is slightly outranked by one or more of the other metabolites in the pathway. The most challenging compounds with respect to specificity are the smaller ones, i.e., those created after the aldolase, presumably because they can easily fit into the larger binding sites, and in some cases are similar to portions of the larger substrates. These "failures" in predicting specificity among these closely related metabolites may represent limitations of the scoring function, or it may be that some of these compounds do act as competitive inhibitors of other enzymes in the pathway.

Our results indicate that docking combined with molecular mechanics rescoring methods do in fact succeed in identifying substrates as top-ranked metabolites, and in general succeed at identifying the correct substrate for the correct enzyme within the pathway. As in our prior work, using a molecular mechanics based scoring function in conjunction with an implicit solvent model improves the results significantly, relative to using a commonly used empirical docking scoring function, which we attribute in part to the highly polar nature of the binding sites. These results are significant because they indicate both that the computational methods are up to the task, and that predicted relative binding affinities can be sufficient to at least exclude a large fraction of the metabolome.

It should also be noted that, although it is encouraging that the cognate ligands rank so highly, there are still quite a few metabolites that outrank them in most cases. We assume (but cannot prove) that most of these do not in fact bind strongly. These presumed false positives are likely due to known limitations of the scoring function, such as treating water as dielectric continuum and neglect of ligand and protein entropy losses. As a practical matter, many of the false positives could be eliminated as potential substrates based on other criteria, such as chemical plausibility, given the known reactions catalyzed by the enzyme superfamily. The binding poses can also be examined to eliminate false positives, as we have done in a recent study where we used an *apo* crystal structure determined by a structural genomics consortium (11). Of course, experimental testing will remain necessary to definitively establish enzymatic function.

These results, as with our prior retrospective study on the enolase superfamily (3), do not directly assess the ability to use these methods in practice to assign functions to uncharacterized enzymes. Here we used crystal structures—or homology models based on crystal structures—that contain at least critical co-factors and in some cases also a product or substrate/intermediate analog. Thus, these results represent the best-case scenario, and are primarily a test of the scoring function and underlying assumptions. Nonetheless, it is noteworthy that 8 of the 10 cases we present here are based on docking to homology models, based on templates with 40–

64% sequence identity. Additionally, as we have previously shown in several cases, the methods used here can still be used productively in “real” applications, using *apo* structures and homology models, to help assign enzyme functions. The primary additional challenge in such applications is predicting conformational changes due to ligand binding (17).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant GM071790. MPJ is a consultant to Schrödinger LLC. We thank Prof. John Gerlt (UIUC) for many helpful conversations.

References

1. Whisstock JC, Lesk AM. *Quart Rev Biophys* 2003;36:307–340.
2. Burkhard R. *J Mol Biology* 2002;318:595–608.
3. Kalyanaraman C, Bernacki K, Jacobson MP. *Biochemistry* 2005;44:2059–2071. [PubMed: 15697231]
4. Tyagi S, Pleiss J. *J Biotech* 2006;124:109–116.
5. Herman JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK. *J Am Chem Soc* 2006;128:15882–15891. [PubMed: 17147401]
6. Herman JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. *Nature* 2007;448:775–779. [PubMed: 17603473]
7. Xiang DF, Kolb P, Federov AA, Meier MM, Federov LV, Nguyen TT, Sterner R, Almo SC, Shoichet BK, Raushel FM. *Biochemistry* 2009;48:2237–47. [PubMed: 19159332]
8. Macchiarulo A, Nobeli I, Thornton JM. Ligand selectivity and competition between enzymes in silico. *Nat Biotechnol* 2004;22:1039–1045. [PubMed: 15286657]
9. Favia AD, Nobeli I, Glaser F, Thornton JM. *J Mol Biol* 2008;375:855–874. [PubMed: 18036612]
10. Song L, Kalyanaraman C, Fedorov AA, Federov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA. *Nat Chem Biol* 2007;3:486–491. [PubMed: 17603539]
11. Kalyanaraman C, Imker HJ, Federov AA, Federov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP. *Structure* 2008;16:1668–1677. [PubMed: 19000819]
12. Rakus JF, Kalyanaraman C, Federov AA, Federov EV, Mills-Groninger FP, Toro R, Bonanno J, Bain K, Sauder M, Burley SK, Almo SC, Jacobson MP, Gerlt JA. *Biochemistry* 2009;48:11546–11558. [PubMed: 19883118]
13. Babbitt PC, Hasson MS, Wedekind JE, Palmer DRJ, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA. *Biochemistry* 1996;35:16489–16501. [PubMed: 8987982]
14. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. *J Med Chem* 2004;47:1739–1749. [PubMed: 15027865]
15. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. *Nucleic Acids Res* 2002;30:402–404. [PubMed: 11752349]
16. Schray KJ, Benkovic SJ, Benkovic PA, Rose IA. *J Biol Chem* 1973;248:2219–2224. [PubMed: 4570471]
17. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. *J Med Chem* 2006;49:534–553. [PubMed: 16420040]

Table 1

Enzymes and substrates in the *E. coli* glycolysis pathway used for testing metabolite docking.

Step	Enzyme	Substrate	SeqID ^a
I	Glucokinase	Glucose	100%
II	Phosphoglucose isomerase	Glucose-6-phosphate	64%
III	Phosphofructokinase	β -D-Fructose-6-phosphate	100%
IV	Fructose biphosphate aldolase	β -D-Fructose-1,6-bisphosphate	40%
V	Triosephosphate isomerase	Dihydroxyacetone phosphate	45%
VI	Glyceraldehyde 3-phosphate dehydrogenase	D-Glyceraldehyde 3-phosphate	58%
VII	Phosphoglycerate kinase	1,3-Bisphospho-D-glycerate	45%
VIII	Phosphoglycerate mutase	3-Phospho-D-glycerate	49%
IX	Enolase	2-Phospho-D-glycerate	51%
X	Pyruvate kinase	Phosphoenol-pyruvate	46%

^aSequence identity between the *E. coli* enzyme and the structural template used to create the homology model of the enzyme. 100% indicates that a crystal structure was used. Further details are provided in Table S1.

Table 2

Ranks (in percent) of metabolites in the glycolysis pathway after docking against each enzyme active site

Ligands											
	Step	I	II	III	IV	V	VI	VII	VIII	IX	X
Enzymes	I	2.3^a	<u>6.4^b</u>	8.0	17.8	40.7	32.0	48.7	9.5	39.1	42.0
	II	12.3	2.7	<u>8.5</u>	7.7	11.3	6.4	7.1	4.3	5.2	4.9
	III	20.1	1.6	3.3	<u>0.01</u>	5.3	5.7	0.4	6.8	5.4	8.3
	IV	22.4	3.6	7.3	3.5	<u>17.2</u>	12.5	2.3	0.2	1.1	0.8
	V	21.7	1.4	3.0	2.0	0.9	<u>0.6</u>	0.3	0.6	1.3	0.8
	VI	23.1	3.6	10.9	5.8	17.4	12.5	<u>7.9</u>	10.8	9.3	14.7
	VII	22.4	3.6	12.4	0.06	11.3	7.3	0.8	<u>8.5</u>	8.8	8.9
	VIII	6.4	<u>-^c</u>	-	-	1.0	0.3	-	0.3	<u>0.5</u>	0.8
	IX	7.5	-	-	-	1.0	0.5	-	3.7	3.3	<u>4.2</u>
	X	14.0	6.2	8.5	16.8	26.4	10.4	14.3	1.7	7.8	6.7

^aDiagonal elements highlighted in bold represent cognate enzyme substrate pairs.

^bUnderlined elements represent enzyme products.

^cLigands rejected by the docking program due to poor scoring are represented by the symbol ‘-’.

Table 3

Ranks (in percent) of metabolites in the glycolysis pathway after MM-GBSA rescoring

Ligands											
	Step	I	II	III	IV	V	VI	VII	VIII	IX	X
Enzymes	I	0.06^a	<u>15.3^b</u>	13.7	45.9	21.2	18.2	47.7	29.8	39.3	41.5
	II	0.8	0.9	<u>0.6</u>	0.01	7.2	10.9	35.9	10.7	13.9	23.4
	III	9.0	15.3	0.1	<u>0.3</u>	10.2	8.0	5.0	63.2	61.1	85.8
	IV	22.1	0.9	0.1	0.3	<u>1.2</u>	4.1	2.5	0.2	1.4	5.0
	V	26.6	2.3	29.3	18.0	0.07	<u>0.06</u>	26.1	16.5	11.1	45.6
	VI	13.4	0.5	0.1	48.3	0.9	0.8	<u>81.7</u>	75.4	29.6	46.6
	VII	16.6	2.7	6.7	0.06	0.4	0.4	0.5	<u>0.5</u>	3.6	2.0
	VIII	8.3	<u>_c</u>	-	-	0.08	0.1	-	0.03	<u>0.01</u>	0.08
	IX	11.4	-	-	-	0.4	0.7	-	0.9	0.2	<u>0.5</u>
	X	21.4	1.1	3.2	6.8	5.7	3.7	0.09	2.3	0.1	1.0

^aDiagonal elements highlighted in bold represent cognate enzyme substrate pairs.^bUnderlined elements represent enzyme products.^cLigands rejected by the docking program due to poor scoring are represented by the symbol '—'.