SCHOOL OF PHARMACY
DEPARTMENT OF PHARMACEUTICAL CHEMISTRY
COMPUTER GRAPHICS LABORATORY
PHONE: (415) 476-2299
E-MAIL: tef@cgl.ucsf.edu

SAN FRANCISCO, CALIFORNIA 94143-0446
FAX: (415) 502-1755

DOE/ER/62269--T1
June 24, 1997

Technical Progress Report

Project Title: Conscription of Proteins for New Functionality
Principal Investigator: Thomas E. Ferrin
Reporting Period: 10/1/96 - 6/1/97 (8 months)
Recipient Organization: University of California, San Francisco
DOE Award Number: DE-FG03-96ER62269
Projected Fund Balance: \$34,416 (17.2% of original award)

A. SCIENTIFIC PROGRESS

MASTER

Aim 1.1: Development of Function-Based Screening Routines

Initial investigation of this problem has centered around using Bayesian Belief Networks (BBN) to evaluate membership in a superfamily. A graduate student working with Prof. Babbitt (Scott Pegg) has done initial evaluation of the methodology for this purpose. That analysis suggests that functional information, including chemical reaction that a substrate can undergo, overall reaction associated with an enzyme, etc. can be used as nodes in the network. This information will supplement the use of structural information, including database search scores, Poisson probabilities, and congruence information from multiple database searches (see Aim 1.2) that will occupy additional BBN nodes. The prototype network will be completed by the end of the summer and we will begin testing its capabilities using known superfamilies [1, 2].

Aim 1.2: Generate Methods for Sorting Output of Database Searches

An algorithm, "re-blast," has been developed that performs multiple BLAST searches [3] using the members of 5 different enzyme superfamilies as query sequences. The output of the searches is then analyzed for congruent hits across 2 or more query sequences. The BLAST searches are set up to save more than 500 of the top scores and the output contains a large majority of scores that are not statistically significant. Congruent hits are sorted by the number of query sequences that "found" a given hit, along with associated scores. The results of the prototype algorithm show that by using a divergent set of previously identified superfamily members as query sequences, all of the known superfamily members are easily identified with very few false positives. Sequences scoring with very poor statistical significance

1. P.C. Babbitt, M. Hasson, J.E. Wedekind, D.J. Palmer, M.A. Lies, G.H. Reed, I. Rayment, D. Ringe, G.L. Kenyon, and J.A. Gerlt, "The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids," *Biochem.* **35**: 16489-16501 (1996).
2. P.C. Babbitt and J.A. Gerlt "Understanding enzyme superfamilies: Chemistry as the fundamental determinant in the evolution of new catalytic activities," *Jour. Biol. Chem.* in press (1997).
3. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, **215**: 403-410 (1990).

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

pgs

27

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

were also readily identified as superfamily members in the output. Most exciting, several new superfamily members were identified that had not been previously found. Again, these new superfamily members were found by several query sequences, but all at very poor statistical significance. This suggests that this approach is capable of mining protein sequence databases for real relationships that would otherwise be too distantly related to give statistically significant scores. Given pre-sorting of superfamily members to generate a set of divergent superfamily members for the initial analysis, the algorithm successfully automates analysis of database output to highly enrich identification of true family members.

The results from this work was presented at an invited talk by Prof. Babbitt at the 6th Annual Conference on Bioinformatics and Genome Research sponsored by the Cambridge Healthtech Institute and held in San Francisco June 11-12, 1997. Future development of the algorithm seeks to automate generation of the initial divergent set of superfamily members to be used as query sequences. This will be done through interactive searches beginning with one sequence and filtering the output to identify candidate query sequences for subsequent searches. The goal is to provide an automated method for identifying distantly related members of a superfamily starting with a single sequence.

Aim 2: Description of Structural Relationships Among Members of a Superfamily

This project has been deferred pending the hiring of a postdoctoral researcher (see Administrative Summary below).

Aim 3: Develop Strategies for Determining Chemically Relevant Three-Dimensional Scaffolds

Before one can accurately predict determinants across a superfamily, it is necessary to establish specificity for individual enzymes using computational tools. Specifically, it is not clear how successful docking methods that are designed for inhibitor design will work for substrate prediction.

Our preliminary results have included the individual dockings using the program Hammerhead [4] of (R) and (S) mandelate, (R)-alpha-phenylglycidate, 150 commercially available alpha-hydroxy carboxylic acids, 450 natural and synthetic amino acids, and a random subset of 1000 compounds from the Available Chemicals Directory to serve as a control group. Results of these dockings have predicted the binding of (R) and (S) mandelate, (R)-alpha-phenylglycidate, 28 alpha-hydroxy carboxylic acids, 101 natural and synthetic amino acids, and 15 compounds out of the control group.

Interestingly, the (R) and (S) isomers of mandelate each form hydrogen bonds with the active site residues shown to be critical in the racemization. However, they are docked in reverse order of what is expected experimentally [5, 6]. Currently, we are using the UCSF DOCK program to investigate this phenomenon to ascertain whether it is directly related to the initial method used (Hammerhead) or whether when one is docking a compound into a site, that the current methods do not model covalent bond formations as in the case of mandelate racemase.

Aim 4: Integration of Computational Methodologies into User Friendly Tools

One of the key procedures in analyzing sequences from a superfamily of proteins is to perform a multiple sequence alignment. These procedures superimpose a set of sequences in order to investigate the similarity between a set of sequences. Because multiple sequence alignment algorithms can only produce approximate alignments, a multiple sequence alignment editor is needed to evaluate and correct the results.

4. W. Welch, J. Ruppert and A.N. Jain, "Hammerhead: Fast, Fully Automated Docking of Flexible Ligands to Protein Binding Sites," *Chemistry and Biology*, **3**: 449-462 (1996).

5. V.M. Power, C.W. Koo, G.L. Kenyon, J.A. Gerlt and J.W. Kozarich, "Mechanism of the Reaction Catalyzed by Mandelate Racemase: 1. Chemical and Kinetic Evidence for a Two-Base Mechanism," *Biochemistry* **30**: 9255-9263 (1991).

6. J.A. Landro, A.T. Kallarakal, S.C. Ransom, J.A. Gerlt, J.W. Kozarich, D. J. Neidhardt, and G. L. Kenyon, "Mechanism of the Reaction Catalyzed by Mandelate Racemase: 3. Asymmetry in Reactions Catalyzed by the H297N Mutant," *Biochemistry* **30**: 9274-9281 (1991).

We are in the process of implementing a multiple sequence alignment editor using the Python programming language [7] and the Tk user interface toolkit [8]. This sequence editor will allow the user not only to analyze multiple sequence alignments, but also to relate them to the three-dimensional structure of the protein (where available) using Chimera molecular modeling system [9]. As part of this project, we did a careful evaluation of other available software, including the Genetics Data Environment (GDE) and LOOK [10] packages. Because neither of these packages can be easily customized to fit into a flexible modeling environment, we reached the conclusion that it was necessary to write our own multiple sequence editor, modeled after the GDE package and implemented in Python. This will allow us to make full use of the interactive three-dimensional molecular graphics capabilities found in Chimera and will provide a base upon which we can easily add new functionality in the future. Although implementation of our multiple sequence editor is not yet complete (and we've not yet chosen a name for it), several features are already operational, including:

- 1) Editing multiple sequence alignments;
- 2) Searching by regular expression using Python's regular expression module;
- 3) Color coding of sequences (text) using Python's regular expression module;
- 4) Color coding of background regions of the alignment.

We expect to complete the implementation of this program during the summer.

A second significant computational tool that is under development is a multiple structure superposition procedure, described by Diamond [11]. This program will enable us to optimally align a series of structures if we can define the atomic correspondence among the structures. The aligned structures may then be used as a basis for constructing a three-dimensional scaffold. The preliminary implementation, using the Python programming language, of this multiple superposition application has been completed and tested. We will be adding a graphical user interface during the summer.

B. ADMINISTRATIVE SUMMARY

Less funds have been expended than anticipated due to our difficulty in hiring a postdoctoral researcher to work on this project. Although we had tentatively recruited Mr. Peter Spiro, who will graduate soon from the Department of Mathematics, Univ. of Utah, he has recently decided to work for a commercial software and database development company (Incyte Pharmaceuticals, Palo Alto, CA) at twice the initial salary we offered. We have since intensified our recruitment activities and now hope to have someone on board within the next month or two. The reality is that the job market for individuals with a background in bioinformatics, as is required for this project, is very much a "buyers market," with far more jobs available than individuals to fill them. We are at a disadvantage in that the salary level we can offer for postdoctoral researcher positions is not competitive with industry. Our intent is to carry forward our unexpended salary funds into year 02 and use these to pay the salary for the postdoctoral researcher, possibly at a higher level than originally budgeted if this proves necessary in order to retain the best qualified person.

Three graduate students are currently participating on this project: Mr. Scott Pegg (1st year student, Pharmaceutical Chemistry), Dr. Manish Butte, MD (2nd year student, Biophysics), and Mr. Bobby Otilar (3rd year student, Biophysics). This level of student involvement is consistent with the training goals of

7. M. Lutz, "Programming Python," O'Reilly and Associates, 1996.

8. J.K. Ousterhout, "Tcl and the Tk Toolkit," Addison-Wesley, 1994.

9. C.C. Huang, G.S. Couch, E.F. Pettersen, and T.E. Ferrin, "Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components," *Pacific Symposium on Biocomputing '96*, World Scientific Publishing, 1996.

10. "Look User's Guide," Molecular Applications Group, Palo Alto, 1996.

11. R. Diamond, "On the multiple simultaneous superposition of molecular structures by rigid body transformations," *Protein Science*, 1: 1279-1287 (1992).

the DOE for the project. Of these, only Mr. Pegg receives (partial) support from this grant; the remaining student stipend support comes from training grants in Biophysics and Pharmaceutical Chemistry.

The original budget included 20% clerical support. We elected not to expend funds for this purpose at this time, since the original justification for this partial position was in support of software dissemination activities and we are not yet at the stage of the project where this activity is warranted. Instead these funds are earmarked for the support of the postdoctoral researcher that is discussed above.

Other than as noted above, expenditures have been as was originally budgeted in our revised project proposal.

Respectively submitted,

/s/

Thomas E. Ferrin, Ph.D.
Professor of Pharmaceutical Chemistry
Director, Computer Graphics Laboratory

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.