

Global Grid Computing at NYU

The Human Proteome Folding Project

By Richard Bonneau
bonneau@nyu.edu

Grid computing can mean different things to different people. For the purposes of this article, we'll define grid computing as the use of the collective processing resources of a network of loosely coupled computers to solve large-scale computational problems. The computers in question are not primarily devoted to the grid: they can be anywhere in the world, can belong to anyone (so long as they choose to participate in the grid), and need only communicate with a central server for brief bursts each time a large part of the computational problem has been completed.

This type of grid strategy has resulted in several projects where people throughout the world can devote their computers' processing power to public research efforts during the times when their computers are not in use. One of the best known of these efforts is *seti@home* (<http://seti.org/>), a public project dedicated to finding signs of extraterrestrial activity in interstellar radio signals. If the input/output (I/O) needs of the computational problem are small and there is no need for careful timing, these grids can scale to include many hundreds of thousands of computers (as in the case of the project described below) to millions (as in the largest projects such as *seti@home*). Given that, on average, less than 10% of an office computer's daily processing capacity is used, these grid computing projects are tapping a valuable and

underused global resource, enabling much larger scale projects than could otherwise be completed.

This article describes the Human Proteome Folding Project, a grid computing effort currently being carried out by NYU in collaboration with IBM (www.grid.org/projects/hpf/). We're using the computing power of millions of computers to predict the shapes of human proteins about which researchers currently know little. From these detailed shapes, we hope to learn about the functions of these proteins, as the shape of a protein is inherently related to how it functions in our bodies.

The resultant database of protein structures and putative functions will let scientists take the next steps towards understanding how diseases that involve these proteins work. We hope that our work on this project will contribute critical public infrastructure for the biological and biomedical community. Only with the amount of computing power available through the World Community Grid (described below) could we hope to complete this project. The scale of the Grid also allows for better sampling, enabling improved accuracy and gene coverage. I'll describe the project in greater detail below, and then go on to explain how you (via your computer) can participate.

PROTEINS

Proteins are the most important molecules in living beings. Just about

everything in your body involves or is made out of proteins. They are structural molecules made up of long chains of smaller molecules called amino acids that act as enzymes and important carriers of biological signals. There are 20 amino acids that combine in different ways to make up proteins. As the chain of amino acids is built, the chain folds (like balling up string) into a more compact mass, ending up in a particular shape. This process is called "protein folding." Many of the things that happen in cells are specifically controlled by protein shapes and the functions conferred by those shapes. For example, a protein in a virus or bacterium may have a particular shape that interacts with human proteins or human cell membrane, enabling it to infect cells. This is an oversimplification, but nevertheless, knowing these shapes helps us gain insight into the biology of disease by understanding protein function.

How proteins fold has been a compelling theoretical and experimental research focus since the early 1950s, when Dr. Christian Anfinsen won the Nobel Prize for showing that proteins fold reproducibly. Most proteins fold to a tight ensemble of possible shapes. These ensembles are amazingly reproducible when compared to other polymers, the discovery of which spawned the new field of polymer physics called protein folding. Eventually, methods for folding proteins computationally

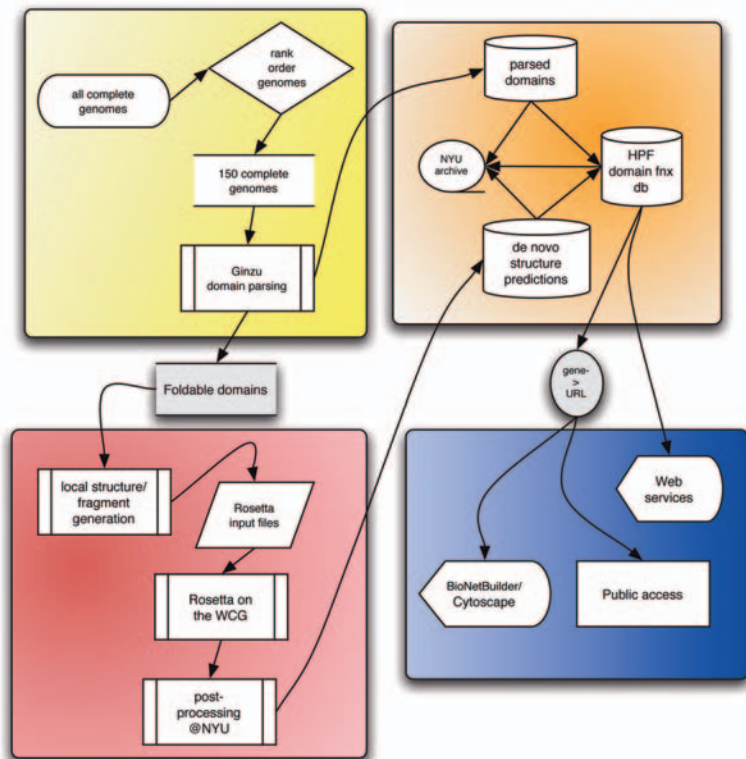


Figure 1. Overall workflow of the Human Proteome Folding Project. Steps carried out on the Grid appear in the block at the bottom left.

came on the scene, one of the more successful of them being the Rosetta software program. Less than five years ago, we (the Rosetta developers community) reached a critical milestone, showing that, for small proteins, we can predict structure well enough to predict aspects of function. Ever since, we have been working hard to improve these methods of extracting function from structure prediction, but the computer power needed for these calculations has been a constant limiting factor.

AN EXPANDING PROTEIN UNIVERSE

In recent years, scientists have sequenced many genomes, including the human genome. Between 20,000 and 30,000 genes found in the human genome encode proteins. The collection of all the human proteins in the genome is known as “the human proteome.” It is not, however, a trivial task to determine from the sequence

of genes their final 3D shape. Protein structure can help us divide proteins into functional classes and thus aid in our organization and understanding of the protein universe. A protein’s shape also gives us clues as to the protein’s function, but there are tremendous barriers to getting structure for whole proteomes.

PROTEIN STRUCTURE PREDICTION

We’ve known for a while that we can get good enough structure predictions using Rosetta to predict function for large numbers of proteins. For example, as part of my PhD at the University of Washington (Baker Lab), I focused on improving the Rosetta method and then on predicting the structures of only 500 key proteins. I was terribly limited by available computer resources, however, and predicting the structure of just these 500 proteins made me somewhat unpopular, as my calcu-

lations monopolized the Baker Lab computer clusters.

What we’d ideally like to do in the Human Proteome Folding Project is attempt to predict the shape of all proteins of unknown function. Due to the massive scale of such a project, this is not possible using existing supercomputing centers or clusters here at NYU, or anywhere else that I know of. One of the features of the problem, however, is that computations on different proteins can be run completely separately, and, in fact, single proteins can be separated into hundreds of smaller independent jobs (a so-called “embarrassingly parallel problem”).¹ Although the supercomputing resources here at NYU are impressive,² they would not be the most cost-effective solution, since parallel problems like ours do not require shared memory or tightly coupled I/O. An attractive aspect of our problem is that data sizes can be kept very small, with large data volume steps separated from processing-intensive steps.

WORLD COMMUNITY GRID PILOT PROJECT

The World Community Grid was started by IBM as a public supercomputing resource; all results are made public, and research efforts are carried out on the non-IBM side by non-profit entities. In 2004, IBM was looking for a biological application they could use to help test their grid (also with the aim that work on the Grid would benefit humanity and help generate positive publicity). Given the inherent grid-ability of our project, we were a natural fit for the World Community Grid. We proposed the Human Proteome Folding Project, and were accepted as the World Community Grid’s pilot project.

Initial work consisted of defining the project’s scope by selecting which proteins were to be folded. In the initial phase of the Human Proteome

1. Wikipedia, http://en.wikipedia.org/wiki/Embarrassingly_parallel

2. See the following *Connect Magazine* articles for more about the Max supercomputer: www.nyu.edu/its/pubs/connect/fall05/ackerman_supercomputer.html and www.nyu.edu/its/pubs/connect/spring06/allison_max.html.

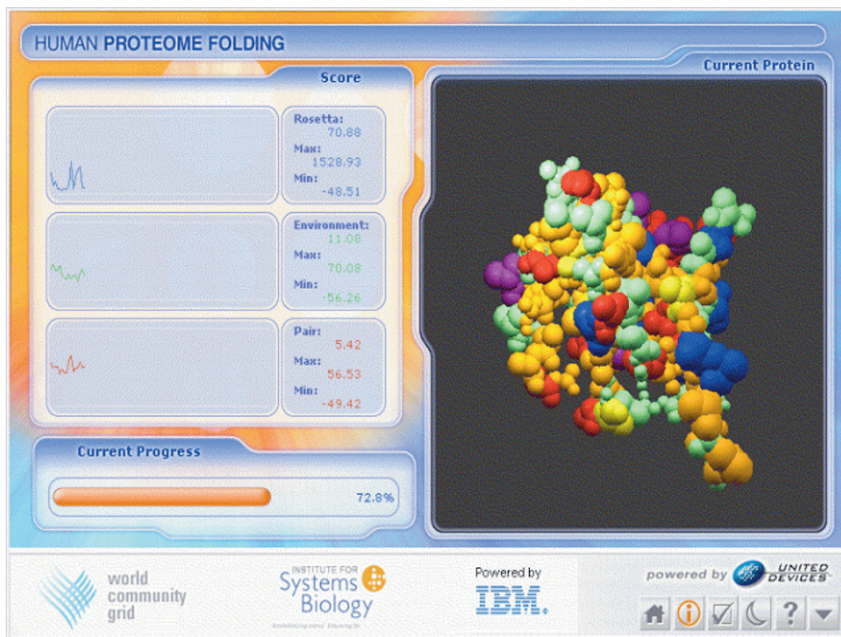


Figure 2. The HPF, Phase 1 client. Upper panels show the structure and simulation scores, as well as progress. The lower bar provides tools for management of the client and its priorities on the volunteer computer.

Folding Project (HPF1), we folded 150,000 proteins of unknown function. In the current, second phase (HPF2), we're focusing on approximately 10,000 proteins (150 genomes) of high interest, including human secreted proteins, cancer biomarkers, and small secreted plasmid proteins. Another task was to construct the Rosetta grid-client. This consisted of changing our code to match special grid-client library calls and testing the server-side pipeline (scripts for breaking work into smaller pieces and recollecting the finished work units). Rick Alther and Viktors Berstis were instrumental in the completion of this work. Robin Wilner and Bill Bovermann led the overall coordination of all the other aspects of the Grid beyond the science, such as the public relations efforts needed to persuade people to download the client.

THE WORKFLOW

The overall workflow of the Human Proteome Folding Project is shown in figure 1 (p. 3). The preprocessing steps consist of the sequence analysis

needed for making the work units. We first exhaust all means of finding sequence matches (the most common way of annotating proteins), and only when we have confirmed that a protein, or a part of a protein, has no matches to known structures do we send it out to the Grid to be folded. This process helps us conserve

resources; the Grid is large, but we still need to maximize the utility of the 500,000 CPU years³ it will give us.

These preprocessing steps are shown in the box at the top left of figure 1. The box at the lower left shows the most CPU-intensive part of the calculation, and the heart of the calculation performed on the Grid. The rightmost boxes depict our meta-database (structure, sequence, localization, and other annotations integrated for proteins in the 150 genomes analyzed) as well as our planned interface to this meta-database.

The large data volumes of HPF1 and the voracious appetite of the Grid for work units posed a challenge, but we were able to complete HPF1 without any down time, any fake work units, and any data loss on the part of NYU or the Institute of Systems Biology (ISB) in Seattle, due in large part to heroic efforts by NYU, the ISB, and IBM team members.

THE RESULTS OF THE FIRST PHASE

Work on HPF1 is drawing to a close, and results from HPF2 (the current

Continued on p. 6, after inset >>

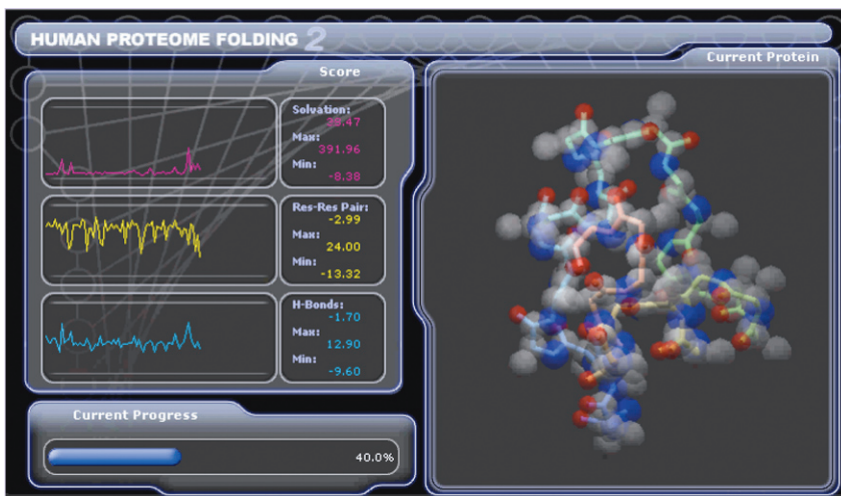


Figure 3. The HPF, Phase 2 client. The panel to the right shows the protein as it folds. The left panels show three components of the score that Rosetta uses to judge protein structures and ultimately return the best-scoring structures to the central server.

3. The CPU, or Central Processing Unit, is a computer component that interprets instructions and processes data in computer programs. CPU years are a unit of measurement for processing time.



Three Grid Collaborations on Disease Research

Following are descriptions of three of the Bonneau Laboratory's main efforts to transfer the protein information returned from the World Community Grid into the hands of groups working on disease research. In general, the two parts of the Human Proteome Folding Project (HPF1 and HPF2) are providing vital bioinformatic support, helping these groups to understand the structure and function of proteins central to their research efforts. In each case, our lab has tailored tools and databases to suit the specific needs of each research effort, and these collaborations will in turn inform our future development of tools for integration of organism-specific information with our structure-prediction-derived information. For each project below, we explain why we have asked you to join the World Community Grid and help us fold these proteins. We encourage you to explore the websites of each of these three groups to learn more about their efforts.

Malaria Proteins

Collaborators: Patrick Duffy & Paul Shannon
www.sbri.org/research/duffy.asp

The Duffy Lab at the Seattle Biomedical Research Institute in Seattle, Washington aims to create a pregnancy malaria vaccine. In 2003, Dr. Duffy and a consortium of laboratories launched the Pregnancy Malaria Initiative to identify the necessary antigens for a malaria vaccine to protect women during pregnancy. They are using bioinformatics, microarray, and proteomics tools to characterize the distinct features of these parasite proteins and evaluate those that may be developed as pregnancy malaria vaccines. The proteins identified by the consortium are now being assessed by the Human Proteome Folding project (HPF2) in order to understand their function and structure so that vaccine designs can be improved.

Using the paradigm established in their studies of pregnancy malaria, the Duffy Lab has also launched a program to develop vaccines against severe childhood malaria. With support from the Grand Challenges in Global Health (GCGH) Program, an international consortium led by the Duffy Lab is now studying the immune responses that protect African children from severe malaria. African children may only suffer one or two episodes of severe malaria before developing resistance, and earlier studies showed that antibody purified from the serum of immune adult Africans could cure young children with malaria. The consortium is thus identifying parasite proteins (*Plasmodium*) that may be targeted by protective antibodies as a key step in developing vaccines for children. As part of the Human Proteome Folding project, we are working with the Duffy Lab to dramatically improve their ability to annotate many of the proteins they have recently found to be important to *Plasmodium* and its specific interactions with its host.

Human Cancer Biomarkers

Collaborators: Leroy Hood & Nathan Price
www.systemsbiology.org/

Multiple groups at the Institute for Systems Biology in Seattle, Washington are currently involved in a coordinated effort to characterize biomarkers that can be used for early diagnosis and sub-classification of human cancers. In particular, specific efforts are underway in the laboratory of Leroy Hood to find prostate, bladder, and ovarian cancer biomarkers. This project coordinates proteomic, microarray,

pathology, and bioinformatics efforts in an attempt to determine reliable and readily-assayable predictors that can be used as markers for diagnosis and selection among alternate therapeutic/intervention regimes.

The Bonneau Lab and the World Community Grid are involved in the functional annotation of putative proteins and proteins of unknown function found in these studies. To date, several hundred putative biomarkers of unknown function have been prioritized and are being processed, along with the other sets of proteins described in this article, on the World Community Grid. The Bonneau Lab has been applying structure-based annotation to elucidate the structure/function of the putative biomarkers discovered using these genome-wide screens.

Gram-Negative Pathogens

Collaborator: David Goodlett
<http://goodlab.mchem.washington.edu/>

Dave Goodlett's laboratory at the University of Washington in Seattle has used *Francisella tularensis* subspecies *novicida* (strain U112), a mouse pathogen, as a model to study virulence in *Francisella tularensis*, a human pathogen that causes Tularemia, also known as "rabbit fever." Both organisms are extremely virulent to their respective hosts, causing high morbidity/mortality if left untreated by antibiotics. The Bonneau Lab's primary role thus far has been to carry out genome annotation for the most difficult proteins in these Gram-negative pathogens, using our structure-inclusive pipeline.

The Goodlett Lab was able to verify that many proteins in these genomes with no homology to other genomes are actively expressed at different conditions, increasing our interest in applying the folding methods described in the accompanying article to these genomes. In combination with genome annotation, about 80% of genes are predicted to be expressed. Of predicted genes, approximately 30% had no homology to genes encoding proteins of known function, preventing corroboration of their authenticity. However, observation of gene products validated the authenticity of more than 50% of these hypothetical genes, representing 23.2% of all expressed genes; no pseudogenes were observed. Finally, we are using Rosetta de novo, fold recognition, and homology-modeling to predict structure and infer function for many of the genes of unknown function.

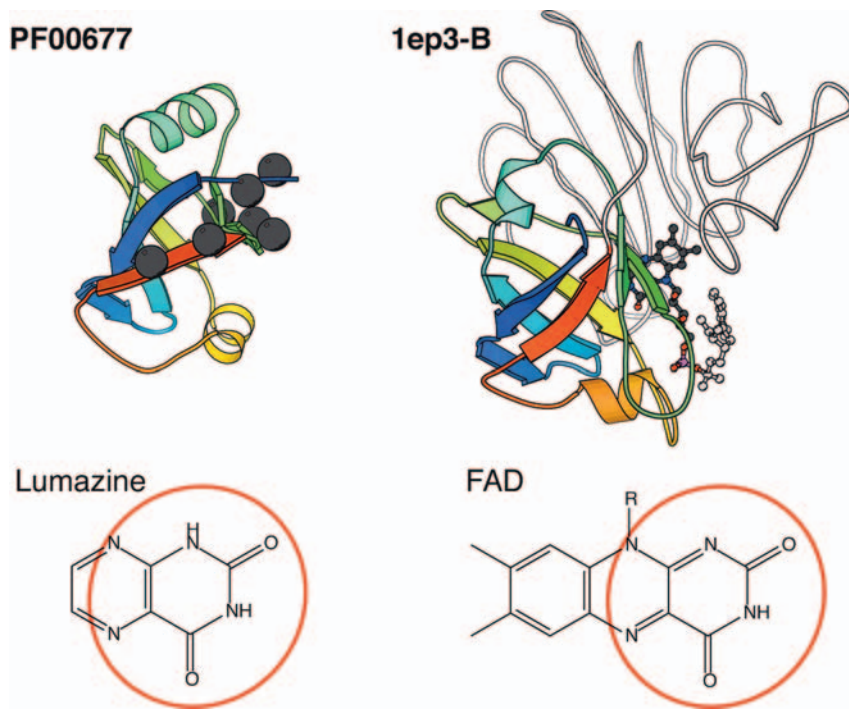


Figure 4. The structure on the left is an example of a Rosetta prediction. The right structure is a structurally similar protein that provides, via structure-structure similarity between the predicted structure (left) and the known structure (right), some clues as to the function of PF00677. (Taken from Bonneau et al., 2001)

>> Continued from p. 4

phase, which I'll describe below) are beginning to arrive at NYU servers. We are now working around the clock—partly due to the fact that the project involves researchers around the globe—to process the results and get them out to the community in our protein domain annotation database.

Alas, predicting the structures is only the first challenge, and current research in my lab centers on integrating the structure-derived data with other sources of protein function (sequence-based, expression and localization measurements, etc.). The beta-release of the database and our release of the results on a model-organism (yeast) are the center of our first paper. Soon after we have debugged the process using the yeast proteome, we'll release the rest of the 150 genomes folded during HPF1 via our meta-database.

HPF2: HIGH-RESOLUTION PROTEIN FOLDING

The second phase of the project (HPF2) is a refinement, using Rosetta in a mode that accounts for greater atomic detail, of the structures resulting from the first phase, HPF1. The project focuses on human secreted proteins, including proteins in the blood, mucus, tears, and the spaces between cells. These proteins can be important for signaling between cells and are often key markers for diagnosis; they have even been found to be useful as drugs, when synthesized and prescribed to people lacking the proteins. HPF2 also focuses on key secreted pathogenic proteins. This phase of the project dovetails with efforts at the Institute for Systems Biology to support predictive, preventive, and personalized medicine, under the assumption that these secreted proteins will be key elements of this medicine of the future.

This second phase of the Human Proteome Folding Project continues where the first left off. The two main objectives are to: 1) obtain higher resolution structures for specific human proteins and pathogen proteins, and 2) further explore the limits of protein structure prediction by further developing Rosetta structure prediction. Thus, the project addresses two very important parallel imperatives, one biological and one biophysical. With the second phase, we are aiming to increase the resolution of a select subset of human proteins. Better resolution is important for a number of applications, including, but not limited to, virtual screening of drug targets with docking procedures and protein design. The second phase of the project will also serve to improve our understanding of the physics of protein structure and help us to further develop our state-of-the-art program, Rosetta.⁴

POTENTIAL FOR FUTURE PROTEIN GRID COMPUTING AT NYU

There are a number of ways we can increase the degree to which spare computer processing cycles are used. Making small private grids—NYU-wide or department-wide, for example—can provide more modest but still quite significant resources for many different analysis pipelines (preprocessing protein sequence for the larger grid, analyzing mass spectrometry data, turning cross links into structures, learning tissue-specific regulatory networks, etc). Another option is for an NYU-wide installation of the World Community Grid client, followed by an NYU-wide attachment to the HPF2 project, a much simpler option, in my opinion. To date, only a small fraction of current global CPU capacity is being used, and grid computing projects have lots of room to grow.

Continued on p. 17 >>

4. We are members of Rosetta Commons, the Rosetta developers' community (www.rosettacommons.org/).

created expectations that extend to radio-like media, such as podcasting. The absence of radio-like production standards in these media is distracting. Next time you listen to an on-location (as opposed to in-studio) interview on NPR, try to pick out where editing has been done. It's usually impossible to tell. Listen for silence between questions: there isn't any. This is because NPR is particularly adept at incorporating ambient sounds into their stories and has been jocularly described as "radio that crunches." While I do not advocate including ambient sound in every interview, podcast producers should keep in mind that audible edits can really distract the audience. It is standard practice to introduce a bit of ambient sound into the silences that are occasionally created by adding pauses during editing. Nothing stands out like dead air.

In addition to ambient sounds, the participants in the interview often breathe audibly. While the listeners may not consciously register these breaths, if one is cut short in an edit, it will become immediately apparent. Indeed, much of the finer work of editing involves adding ambient sound or copying a breath from one portion of an interview

to another. The object is always to make the editing inaudible so that the listener can concentrate on the podcast's content.

EXPANSION & GLOBALIZATION

Six months ago, I was approached by Dr. Kazuo Tsubota, Chairman of the Department of Ophthalmology of Keio University in Tokyo. A faithful listener of *As Seen From Here*, Dr. Tsubota wanted to introduce its contents more widely within the Japanese ophthalmic community. At that time, the audience of *As Seen From Here* was international but concentrated in the English speaking world. Dr. Tsubota's idea was to preface each English language podcast with a Japanese language summary covering the gist of the podcast. His contention was that, while most Japanese ophthalmologists spoke some English, many would feel too intimidated to listen to an entire interview exclusively in English. With a Japanese language summary at the beginning of the program, many would feel comfortable enough to stay on for the English interview.

I saw great value in this approach and have since established collaborations with Peking University, Yonsei

University, and Rajavithi Hospital to produce editions of the podcast with summaries in Mandarin, Korean, and Thai for release this fall. In addition, based at least in part on the success of *As Seen From Here*, the NYU School of Medicine has had the foresight to create an office to take full advantage of podcasting technology. Called The New Media Project, it is, to my knowledge, the first office in any medical school devoted to the production and dissemination of podcasts. Our work includes replicating the ophthalmology podcast in other medical fields and producing both audio and video podcasts for graduate and undergraduate medical education.

The long term effects that the powerful and versatile medium of podcasting will have on the dissemination of scientific, educational, and business material are not yet apparent, but are certain to be substantial.

Joshua Young, M.D., is a practicing ophthalmologist, Director of the New Media Project at the NYU School of Medicine, Producer of As Seen from Here, and Director of Podica (www.podica.com), a podcasting consulting and production service for business.

>> Continued from p. 6

HOW TO PARTICIPATE IN THE GRID

Don't be left out of the satisfying fun that is grid computing! Anyone wishing to participate should visit www.worldcommunitygrid.org/ to download the free software. Thanks to IBM and United Devices, installing and managing the software client is secure and easy. We also have mechanisms for helping whole institutions push out installations; interested parties should contact IBM via the "About Us" form within the "Become a Partner" section of the World Community Grid website.

Once the United Devices or open source BOINC client is installed,

whether on Windows, Macintosh or Linux, the next step is to attach your client to HPF2 by typing in the following project URL when prompted: www.worldcommunitygrid.org/projects_showcase/viewHpf2Research.do. If you wish, you can also join a team by creating your own or selecting an existing one, the New York University team, for example. Another Grid project of note is working to dock small molecules into HIV protein structures in a search for possible new HIV drugs. I encourage anyone reading this article to convince multiple friends to download the client, as well!

For more information about the Human Proteome Folding Project,

select the project name from the list in the "Research" section of the World Community Grid website: www.worldcommunitygrid.org/.

Dr. Richard Bonneau recently joined NYU's Departments of Biology and Computer Science as part of a joint initiative of Computation in Science and Society and NYU's Center for Comparative Functional Genomics. He has played a critical role in the development and deployment of Rosetta, the state-of-the-art protein-folding program described in this article. See <http://cs.nyu.edu/~bonneau/> for more about Dr. Bonneau's research efforts.