Help Conquer Cancer

Welcome to the Help Conquer Cancer Project, and thank you for participating in this important research. The mission of our project is to improve the results of protein X-Ray crystallography, which helps researchers not only annotate unknown parts of the human proteome, but importantly, improves their understanding of cancer initiation, progression and treatment. Although our focus is cancer, resulting from this project will help other diseases as well.

Analyzing the results from this experiment will also lead to better understanding the underlying principles of protein crystallography. For the first time, a comprehensive crystallography image analysis will be done, which was impossible before due to computational complexity. In turn, *CrystalVision* will be improved to provide faster and more accurate image classification. This will lead to more protein structures being determined. In addition, determining more 3D structures of proteins will also lead to improved *in silico* structure prediction.

Thank you again for making vital contribution to the project. It is highly appreciated, as it would take us over 162 years on the largest computer available to us to finish this computation. Although, we expect to finish the project in about 1-2 years on the World Community Grid, we will start analyzing the data incrementally. We will keep you informed about our results.

Mission
The mission of Help Conquer Cancer is to improve the results of protein X-Ray crystallography, which helps researchers not only annotate unknown parts of the human proteome, but importantly improves their understanding of cancer initiation, progression and treatment.

Significance
In order to significantly impact the understanding of cancer and its treatment, novel therapeutic approaches capable of targeting metastatic disease (or cancers spreading to other parts of the body) must not only be discovered, but also diagnostic markers (or indicators of the disease), that can detect early-stage disease, must be identified.

Researchers have been able to make important discoveries when studying multiple human cancers, even when they have limited or no information at all about the involved proteins. However, to better understand and treat cancer, it is important for scientists to discover novel proteins involved in cancer, and their structure and function.

Scientists are especially interested in proteins that may have a functional relationship with cancer. These are proteins that are either over-expressed or repressed in cancers, or proteins that have been modified or mutated in ways that result in structural changes to them.

Improving X-Ray crystallography will enable researchers to determine the structure of many cancer-related proteins faster. This will lead to improving our understanding of the function of these proteins, and enable potential pharmaceutical interventions to treat this deadly disease.

X-Ray Crystallography

One of the favored methods for protein-structure determination is X-ray crystallography. Through this method, scientists use the high-throughput crystallization pipeline to help not only annotate unknown parts of the human proteome, which in turn will help to improve our understanding of cancer initiation, progression and treatment. (NOTE: There are other approaches to understanding the structure and function of proteins, including the Human Proteome Folding Project also running on World Community Grid. Given the essential nature of this work, it's important to advance every research technique to complete our understanding of the human organism and disease.)

There are two main steps involved in X-ray crystallography:

1. Crystallizing the protein. Although a lot more complex, this is similar to putting sugar into a cup of water and letting it sit for a while. Once the water evaporates, tiny sugar crystals appear.
2. Sending X-rays through the crystal and depending on how they diffract, a mathematical model is used to determine and observe the protein's structure.

Crystallizing the protein is not a straightforward procedure. There are many thousands of possible conditions that affect the process (concentration of a protein and solution, temperature, pH, chemical additives, etc.), but scientists must find the appropriate combination of these conditions for a protein to crystallize. For example, with sugar, if you change the water to another liquid, change the temperature or concentrations, you may not get a crystal. Similarly, for a given protein, the challenge is to know what conditions will lead to forming a crystal – what solution, what temperature, pH, etc.

The resultant protein crystal also must be well-formed and large enough in order for x-rays to detect the protein's structure at high resolution. If the conditions are not perfect for crystallizing the protein, the process can result in either a micro-crystal, which is too small for the protein's structure to be determined; a precipitate, which shows some changes, but does not lead to crystallization event directly**;** or no change may have occurred at all.

Frustrating the situation is that, as yet another barrier to progress, usually the more important the protein is to cancer research, the harder that protein is to crystallize. Many proteins involved in cancer are long chains, or they require additional proteins to properly fold and cannot be crystallized by themselves.

In order to run the millions of combinations necessary to successfully crystallize a protein, scientists have used robots to perform the work. Robots are able to put in place the various crystallization conditions faster and more accurately. To further facilitate the process, result of each of the millions of crystallization experiments are photographed.

Currently, scientists at Hauptman-Woodward Medical Research Institute in Buffalo (HWI) have run more than 86 million crystallography experiments for more than 9,400 proteins. As a result, they have 86 million pictures of these proteins that have gone through the X-ray crystallography high-throughput screening pipeline. Each of these pictures needs to be analyzed to determine what the result of the experiment is – i.e., crystal, precipitate, phase separation, skin effect, no change.

One of the challenges is the tremendous size of these datasets, which requires over 25 TB of storage (or equivalent to more than 9,000 DVDs). IBM's Blue Gene supercomputer has provided assistance in this phase of the work, by running a special image compression algorithm to reduce the size of these images, without losing content. The other challenge is to comprehensively analyze an image to determine the crystallization outcome, a task that requires approximately 10 hours to process on a single computer. Researchers would thus require almost 100,000 years to analyze the existing pictures.

World Community Grid and "Help Conquer Cancer"
Using the power of World Community Grid, scientists at the Ontario Cancer Institute, Princess Margaret Hospital, and the University Health Network will process the existing 86 million images of proteins that have been screened in the high-throughput crystallization pipeline at the HWI in Buffalo. World Community Grid will run a *CrystalVision* program that the researchers at OCI have developed to analyze the features of individual images to determine the outcome of the crystallization screen – crystal, micro crystal, phase separation, skin, or a precipitate, or if no change occurred.

If a crystal occurs, crystallographers can put the protein through the optimization process to determine the optimal conditions for the crystallization, and in turn perform a diffraction experiment to determine the structure of the protein. What's more, scientists can compare proteins that have successfully crystallized against proteins of unknown structure that have similar characteristics, based on the results from the crystallization screen. This can be the starting point for crystallization for these proteins so that their structure can be determined.

If the crystal produced was not well-formed or large enough, scientists can still use the information to help them better determine the conditions necessary to create a well-formed crystal. For example, they may learn that Protein X and Condition A resulted in a micro crystal, and Protein A and Condition Z resulted in a micro crystal as well. Based on this information, they can then run additional experiments to deduce what conditions need to be optimized to create a larger and well-formed crystal.

Analyzing the results from this experiment will also lead to better understanding the underlying principles of protein crystallography. For the first time, a comprehensive crystallography image analysis will be done, which was impossible before due to computational complexity. In turn, *CrystalVision* will be improved to provide faster and more accurate image classification.

Improving the protein crystallography pipeline will enable researchers to determine the structure of many cancer-related proteins faster. This will lead to improving our understanding of the function of these proteins, and enable potential pharmaceutical interventions to treat this deadly disease.

Researchers
Igor Jurisica, Principal Investigator, Ontario Cancer Institute

Computational scientists include:
C. Anders Cumbaa, Research Associate, Ontario Cancer Institute

Collaborators
Dr. George DeTitta, Chief Executive Officer, Hauptman-Woodward Medical Research Institute and Chairman, Department of Structural Biology
Joseph R. Luft, Principal Investigator, Hauptman-Woodward Medical Research Institute
Michael Malkowski, Principal Investigator, Hauptman-Woodward Medical Research Institute

Project FAQs

What are the potential benefits of the "Help Conquer Cancer" project?
There are several direct and indirect benefits of the project. For the first time, scientists will execute a comprehensive image analysis and classification of crystallography images. This will lead to better understanding of the crystallization process, and will enable scientists to improve the accuracy and speed of *CrystalVision*. Improved understanding of the crystallization process and improved *CrystalVision* also will enable more disease proteins to be crystallized faster. Finally, more 3D structures will improve our understanding of disease and potentially its treatment, and will lead to improved *in silico* structure prediction.

What computers can run the "Help Conquer Cancer" Project?
Due to the inherent granularity of our image analysis problem, there are very modest memory and CPU requirements for the compute nodes, yet without having access to thousands of CPUs, we would not be able to process 80 million images in a reasonable time. Multiple platforms will be able to run the project; we are launching a Linux and Windows compiled code first, Macintosh OS to follow.

What will World Community Grid's calculations produce?
On the lowest level, *CrystalVision* will compute thousands of image features for each crystallography image. These data objectively measures characteristics of the image, which in turn enable scientists to use a classification system to discern image classification. In turn, this will enable them to automatically and objectively characterize results from the high-throughput crystallization screens. This will then enable us to apply data mining techniques to optimize future crystallization experiments.

What will happen with the data generated by all these calculations?
After careful analysis, evaluation and interpretation of the results, all the results will be published in the public domain. Our first goal is to improve the *CrystalVision* system to enable automated, accurate and fast crystallography image classification. This algorithm will then be deployed at Hauptman-Woodward Medical Research Institute to ensure that this public high-throughput crystallography screening facility will speed up crystallization of many disease-related proteins.

When will this project be completed?
Once the project starts, we will have a better idea about the time required to process the images on World Community Grid. This will be determined by the number of suitable computers and the number of projects being concurrently executed on World Community Grid. However, we have several interesting subsets of images, and those will be analyzed first. Thus, preliminary results will be available after a few weeks. These images comprise set of images previously analyzed by an earlier version of *CrystalVision*, and also by multiple human experts.