Update April 2009

Thank you for your continuing support of the Help Conquer Cancer project (HCC). We appreciate all the computing resources you donate to this and other exciting and useful research at WCG.

Since the launch of HCC project in November 2007, WCG members contributed almost 29,226 years of run time, averaging almost 56 years of computation a day. To date 38,218,562 results were returned (*Statistics Last Updated: 07/04/09 00:05:55*).

Summary

This month marks a milestone in HCC project: 25% of our work units have been processed on the WCG – representing 3 million crystallization trials on over 2,000 proteins. Analysis of these results is proceeding on multiple fronts. Our latest crystal-finding classifier, generated from HCC results, can identify 4 out of 5 images containing protein crystals, greatly reducing the effort of manually searching through images for crystals, and thereby increasing the rate of protein structure solution. Using this classifier on a set of 5.7 million images, we recently identified 11 proteins with favourable crystallization conditions, all of which are homologous to human proteins on our cancer target list (spanning lung, ovarian, prostate and head & neck cancers). Once crystallized, we may be able to characterize function of these cancer targets.

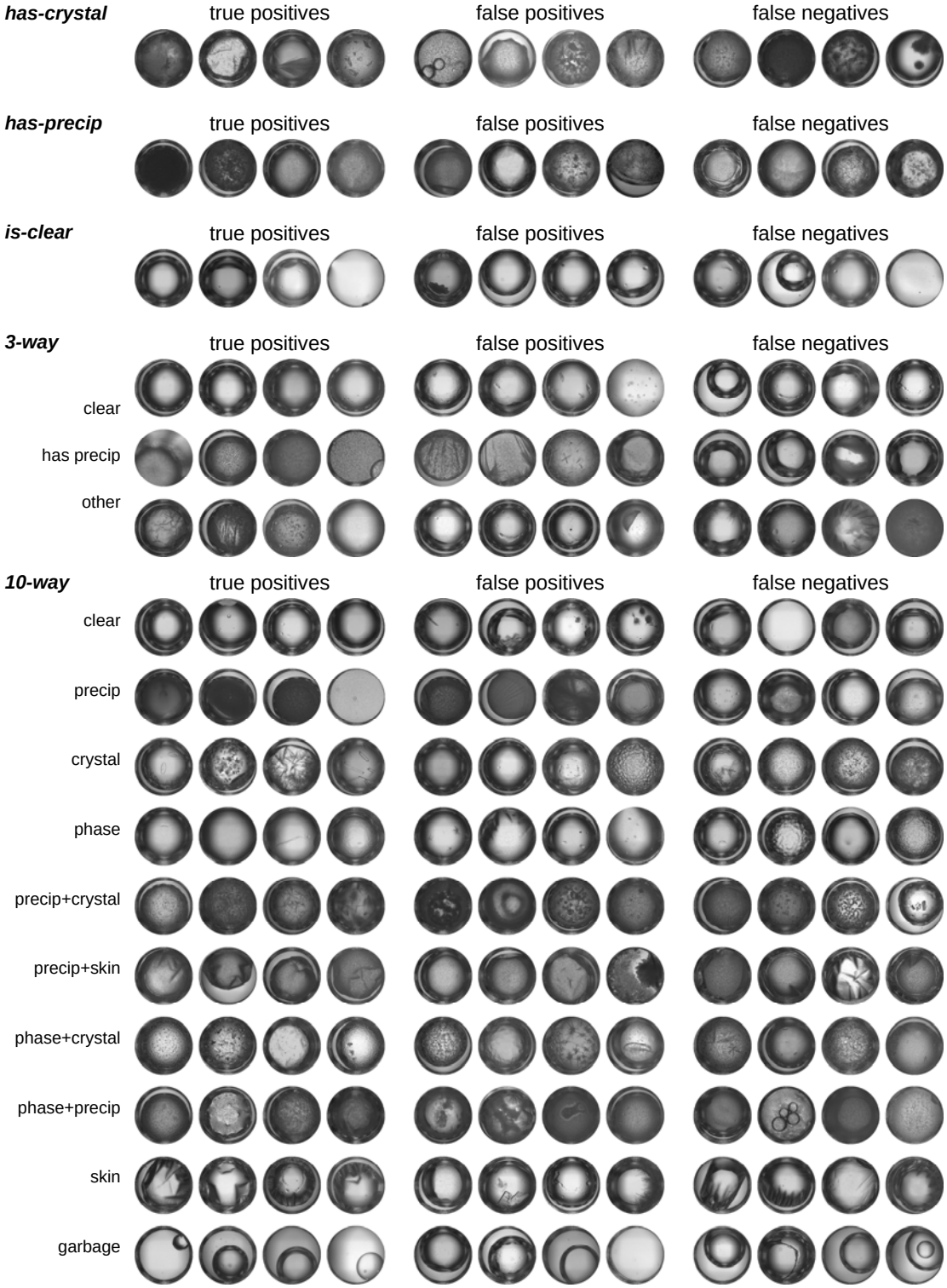
Results

Image classification

We currently use a modified Boosting Feature Selection (BFS) algorithm to objectively and automatically select the most informative features from the total set of 14,908 (all Grid-computed features + 2,533 derived), and generate classifiers. We extended the stock algorithm for multiclass classification tasks, extended the logic to handle probabilistic (rather than yes/no) base learners, and modified the algorithm so that the resulting ensemble classifier was a weighted Naïve Bayes model. We parallelized the algorithm (using MPI calls) for use on our cluster, in order for the training data to fit in (distributed) RAM, and for reduced execution time.

We used BFS to train five classifiers for five different classification goals: 10-way (clear / precip / crystal / phase / skin / garbage / precip & crystal / precip & skin / phase &crystal / phase & precip), 3-way (clear / precipitate / other), is-clear, has-precip, and has-crystal. The has-crystal classifier correctly identifies 4 out of 5 images containing crystals and its ROC score is 85% when multiple time-points are considered together. Results of other classifiers reveal trends: clear drops are detected accurately with 10-20% false-positive rate, 1-5% false-negative rate. Precipitates are recognized with 5% false-positive rate, 6% false-negative rate.

Future work will focus on improving classification of compound categories that include precipitate as they remain difficult to distinguish accurately (e.g., precipitate & crystal vs. precipitate & phase). Figure 1 shows some examples of classification results for individual image classes and specific 3- and 10-way classifiers.



Feature optimization

Assembled together, the ever-growing results generated by the Help Conquer Cancer on the World Community Grid constitute an 18 million × 14,908-element matrix. In an effort to further improve performance of automated image classification, we continue to mine it for meaning and hidden structure, identifying highly-correlated (and therefore redundant) features, and measuring the informative value of image-analysis features and feature clusters on protein crystallization state.

Part of the feature-optimization process involves analyzing the effectiveness of fundamental image-processing parameters. One such parameter is brightness correction.

To determine whether the inherent variation in image brightness levels negatively effects image classification accuracy, we re-computed 11,647 features from all images in our training set, pre-processing each image with a brightness-correction step.

Figure 2 shows the effect of brightness correction on the information content of resulting image features. Features lying above the y=x diagonal show improved information content from brightness correction. Features below the diagonal show diminished information content. Brightness correction boosts information content of image features by 1%. The effect is slightly greater for some high-information-content features.

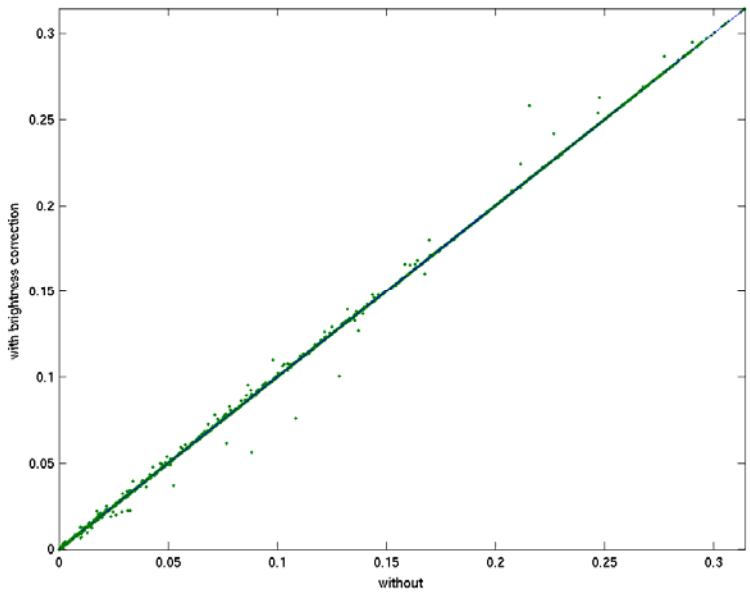


Figure 2 Effect of brightness correction on image-feature information content.

Focusing on cancer targets

To significantly impact cancer research, novel therapeutic approaches for targeting metastatic disease and diagnostic markers reflective of changes associated with early stage disease must be discovered. Better drugs must be rationally designed, and current drugs made more efficacious either by re-engineering or by information-based combination therapy.

Identification and interpretation of cancer biomarkers can be significantly improved by combining multiple types of data, and integrating protein interaction and pathway information. Despite the application of many low- and high-throughput protein interaction detection methods, a large number of false negatives are present in the currently known human interactome (graph where nodes are proteins and edges represent interactions).

To enable systematic analysis of cancer targets, we first integrate multiple gene expression data in lung, ovarian, prostate and head&neck cancers (http://ophid.utoronto.ca/cdip). Second, we expand the current human interactome - 13,509 proteins connected by 100,319 non-redundant interactions (I2D; http://ophid.utoronto.ca/i2d), by using the FPClass association mining algorithm that combines evidence from sequence, gene expression and annotation databases to predict probability of interaction among 74,944 proteins and fragments in SwissProt (http://ca.expasy.org/sprot/). Combined, this resource comprises 256,957 interactions linking 21,543 proteins. The resulting interactome is then visualized and analyzed in NAViGaTOR (http://ophid.utoronto.ca./navigator).

Focusing on frequently and significantly up-regulated targets in the four cancers, linking them on the protein-protein interaction network, and combining with known human protein structures and current PSI targets, this integrative analysis improves characterization and interpretation of putative cancer biomarkers, enables us to verify predicted protein interactions, and implicates new cancer-related targets for PSI pipeline.

Of 5,725,062 images recently processed by the Help Conquer Cancer project, we have identified 11 human homologues with favourable crystallization conditions, all of which are on our cancer target list (spanning lung, ovarian, prostate and head & neck cancers). Once crystallized, we may be able to resolve function of these cancer targets.

Recently published work

Crystallography

1. Snell, E.H., J.R. Luft, S.A. Potter, A.M. Lauricella, S.M. Gulde, M.G. Malkowski, M. Koszelak-Rosenblum, M.I. Said, J.L. Smith, C.K. Veatch, R.J. Collins, G. Franks, M. Thayer, C. Cumbaa, I. Jurisica, and G.T. Detitta, *Establishing a training set through the visual analysis of crystallization trials. Part I:*

approximately 150,000 images. Acta Crystallogr D Biol Crystallogr, 2008. **64**(Pt 11): p. 1123-1130.
2. Snell, E.H., A.M. Lauricella, S.A. Potter, J.R. Luft, S.M. Gulde, R.J. Collins, G. Franks, M.G. Malkowski, C. Cumbaa, I. Jurisica, and G.T. DeTitta, *Establishing a training set through the visual analysis of crystallization trials. Part II: crystal examples.* Acta Crystallogr D Biol Crystallogr, 2008. **64**(Pt 11): p. 1131-1137.

Cancer research

1. Hui, A.B., W. Shi, P.C. Boutros, N. Miller, M. Pintilie, T. Fyles, D. McCready, D. Wong, K. Gerster, I. Jurisica, L.Z. Penn, and F.F. Liu, *Robust global micro-RNA profiling with formalin-fixed paraffin-embedded breast cancer tissues.* Lab Invest, 2009.
2. Dong, J., T. Kislinger, I. Jurisica, and D.A. Wigle, *Lung cancer: developmental networks gone awry?* Cancer Biol Ther, 2009. **8**(4): p. 312-318.
3. Boutros, P.C., S.K. Lau, M. Pintilie, N. Liu, F.A. Shepherd, S.D. Der, M.S. Tsao, L.Z. Penn, and I. Jurisica, *Prognostic gene signatures for non-small-cell lung cancer.* Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2824-2828.
4. Zavareh, R.B., K.S. Lau, R. Hurren, A. Datti, D.J. Ashline, M. Gronda, P. Cheung, C.D. Simpson, W. Liu, A.R. Wasylishen, P.C. Boutros, H. Shi, A. Vengopal, I. Jurisica, L.Z. Penn, V.N. Reinhold, S. Ezzat, J. Wrana, D.R. Rose, H. Schachter, J.W. Dennis, and A.D. Schimmer, *Inhibition of the sodium/potassium ATPase impairs N-glycan expression and function.* Cancer Res, 2008. **68**(16): p. 6688-6697.
5. Tomasini, R., K. Tsuchihara, M. Wilhelm, M. Fujitani, A. Rufini, C.C. Cheung, F. Khan, A. Itie-Youten, A. Wakeham, M.S. Tsao, J.L. Iovanna, J. Squire, I. Jurisica, D. Kaplan, G. Melino, A. Jurisicova, and T.W. Mak, *TAp73 knockout shows genomic instability with infertility and tumor suppressor functions.* Genes Dev, 2008. **22**(19): p. 2677-2691.
6. Sodek, K.L., A.I. Evangelou, A. Ignatchenko, M. Agochiya, T.J. Brown, M.J. Ringuette, I. Jurisica, and T. Kislinger, *Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT).* Mol Biosyst, 2008. **4**(7): p. 762-773.
7. Shedden, K., J.M. Taylor, S.A. Enkemann, M.S. Tsao, T.J. Yeatman, W.L. Gerald, S. Eschrich, I. Jurisica, T.J. Giordano, D.E. Misek, A.C. Chang, C.Q. Zhu, D. Strumpf, S. Hanash, F.A. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V.E. Seshan, M. Meyerson, R. Kuick, K.K. Dobbin, T. Lively, J.W. Jacobson, and D.G. Beer, *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nat Med, 2008. **14**(8): p. 822-827.
8. Ponzielli, R., P.C. Boutros, S. Katz, A. Stojanova, A.P. Hanley, F. Khosravi, C. Bros, I. Jurisica, and L.Z. Penn, *Optimization of experimental design parameters for high-throughput chromatin immunoprecipitation studies.* Nucleic Acids Res, 2008.
9. Niu, Y. and I. Jurisica, *Detecting Protein-Protein Interaction Sentences Using a Mixture Model*, in *Natural Language and Information Systems (NLDB'08), Lecture*

Notes in Computer Science, E. Kapetanios, V. Sugumaran, and M. Spiliopoulou, Editors. 2008, Springer Verlag: Berlin / Heidelberg. p. 352-354.
10. Jurisicova, A., I. Jurisica, and T. Kislinger, *Advances in ovarian cancer proteomics: the quest for biomarkers and improved therapeutic interventions.* Expert Rev Proteomics, 2008. **5**(4): p. 551-560.
11. Gortzak-Uzan, L., A. Ignatchenko, A.I. Evangelou, M. Agochiya, K.A. Brown, P. St Onge, I. Kireeva, G. Schmitt-Ulms, T.J. Brown, J. Murphy, B. Rosen, P. Shaw, I. Jurisica, and T. Kislinger, *A proteome resource of ovarian cancer ascites: integrated proteomic and bioinformatic analyses to identify putative biomarkers.* J Proteome Res, 2008. **7**(1): p. 339-351.
12. Aviel-Ronen, S., B.P. Coe, S.K. Lau, G. da Cunha Santos, C.Q. Zhu, D. Strumpf, I. Jurisica, W.L. Lam, and M.S. Tsao, *Genomic markers for malignant progression in pulmonary adenocarcinoma with bronchioloalveolar features.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10155-10160.

Recent (selected) presentations

- Jurisica, I. PSI target prioritization, cancer target selection and interpretation using protein-protein interaction prediction and analysis. *NIH NIGMS 2009 Workshop on Enabling Technologies for Structural Biology*, Natcher Conference Center, NIH, Washington DC, March 4-6, 2009.
- Cumbaa, C. A. and I. Jurisica. Crystallization image analysis – progress update. *NIH NIGMS 2009 Workshop on Enabling Technologies for Structural Biology*, Natcher Conference Center, NIH, Washington DC, March 4-6, 2009.

Thank you,

C. A. Cumbaa and I. Jurisica