IBM Computational Biology Center, IBM Research
gustavo@us.ibm.com

SEEKING THE WISDOM OF THE CROWDS THROUGH CHALLENGE-BASED COMPETITIONS IN BIOMEDICAL RESEARCH

Outline

- Crowdsourcing and challenges
- Benefits of crowd-sourcing through collaborative-competitions
- The Sage-DREAM Breast Cancer Prognosis Challenge
- The NCI-DREAM Drug Sensitivity Prediction Challenge

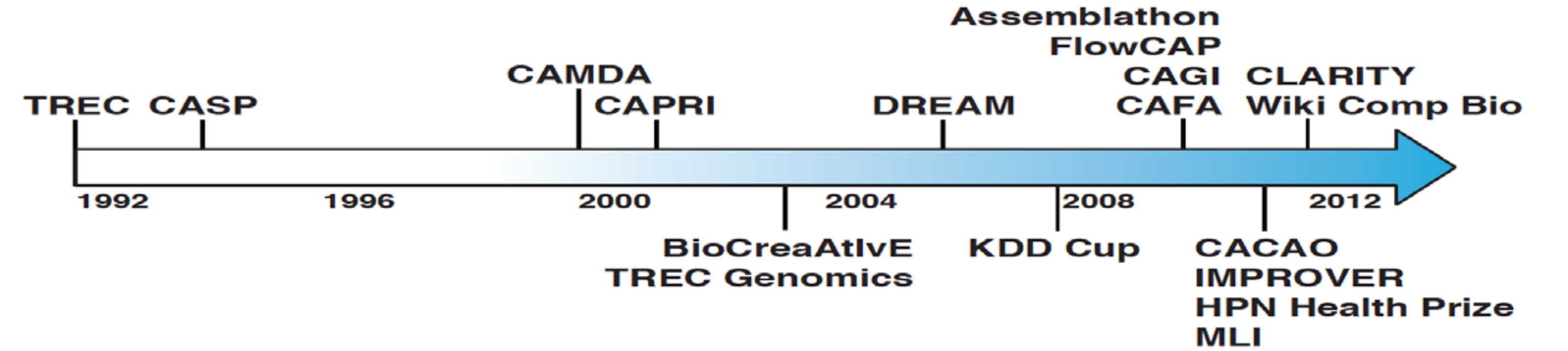
Crowdsourcing and Challenges

Crowdsourcing: The practice of soliciting content, ideas, solutions from a large group of people, especially the online community.

E.g., Protein folding solutions have been generated through a crowdsourcing game: FoldIt.

Challenge: A crowdsourcing based approach to solve a problem

E.g., Dialogue for Reverse Engineering Assessment and Methods (DREAM) challenges in cellular network inference



Benefits of crowdsourcing

- Performance Evaluation
  - Assess whether relevant problems can be addressed computationally: E.g., can drug sensitivity be predicted?
  - Discover the best methods via blind, unbiased, and rigorous method assessment
- Sampling the method space
  - Understand the diversity of methodologies presently being used to solve a problem

Benefits of crowdsourcing

- Community Building
  - Make high quality, well-annotated data accessible.
  - Foster community collaborations on fundamental research questions.
  - Determine robust solutions through community consensus: “The Wisdom of the Crowds.”

The Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge

Goals: Use crowdsourcing to assess whether breast cancer survival can be accurately predicted

Training data set: Genomic and clinical data from 2000 women diagnosed with breast cancer (Metabric data set).

Data access and analyses: Sage Bionetworks' Synapse

Compute resources: Standardized virtual machines for each participant donated by Google

Model scoring: models submitted to Synapse for scoring on a real-time leaderboard

Participation: 1,700 models tested by 48 participating teams, 35 countries

Unique Attributes

- Open source and code-sharing:
  - Standardized computational infrastructure helps participants use code submitted by others in their own models
  - All models’ behavior and performance must be reproducible
- New dataset for final validation to determine winning model:
  - Derived from approx. 200 breast cancer samples
  - Data generation funded by Avon
  - Winning model: the most accurate in predicting survival for independent datasets, following training on the Metabric dataset
- Challenge assisted peer-review
  - Overall winner team can submit a pre-accepted article about their winning model to Science Translational Medicine

NCI-DREAM Summit

- DRUG Challenges and timelines
  - On April 23, 2012 about 20 researchers active on systems pharmacology of cancer gathered at the NCI
  - After a day of discussion and breakout sessions, several possible challenges were suggested
  - In subsequent discussions, based on available blind data, two candidate challenges were selected for refinement.
    - Predicting drug sensitivity in a large collection of BC cell lines
    - Predicting drug synergy in human B cells
  - Challenge data was released in early June 2012, submissions were received in early October, and results were announce in late October

The NCI-DREAM Drug Sensitivity Prediction Challenge

- Goals: Use crowdsourcing to identify computational approaches that best predict therapeutic responses
- Challenges:
  - Sub-challenge 1. Predict sensitivity of 31 compounds in 18 cell lines, given their sensitivity profiles in 35 cell lines and genomic information for all lines
  - Sub-challenge 2. Predict responses to 91 pairwise combinations of 14 compounds in Ly3 human B-cell lymphoma cells
- Data provenance and accessibility:
  - Generated in ongoing ICBP studies but yet unpublished. Data was curated for the challenge and made accessible via the DREAM website upon registration
- Participants:
  - 47 teams and 31 teams participated in sub-challenge 1 and 2, respectively, from more than 30 countries

Best Performers

Sub-Challenge 1:

TeamFIN: Helsinki Institute for Information Technology,
Aalto University, Helsinki Finland

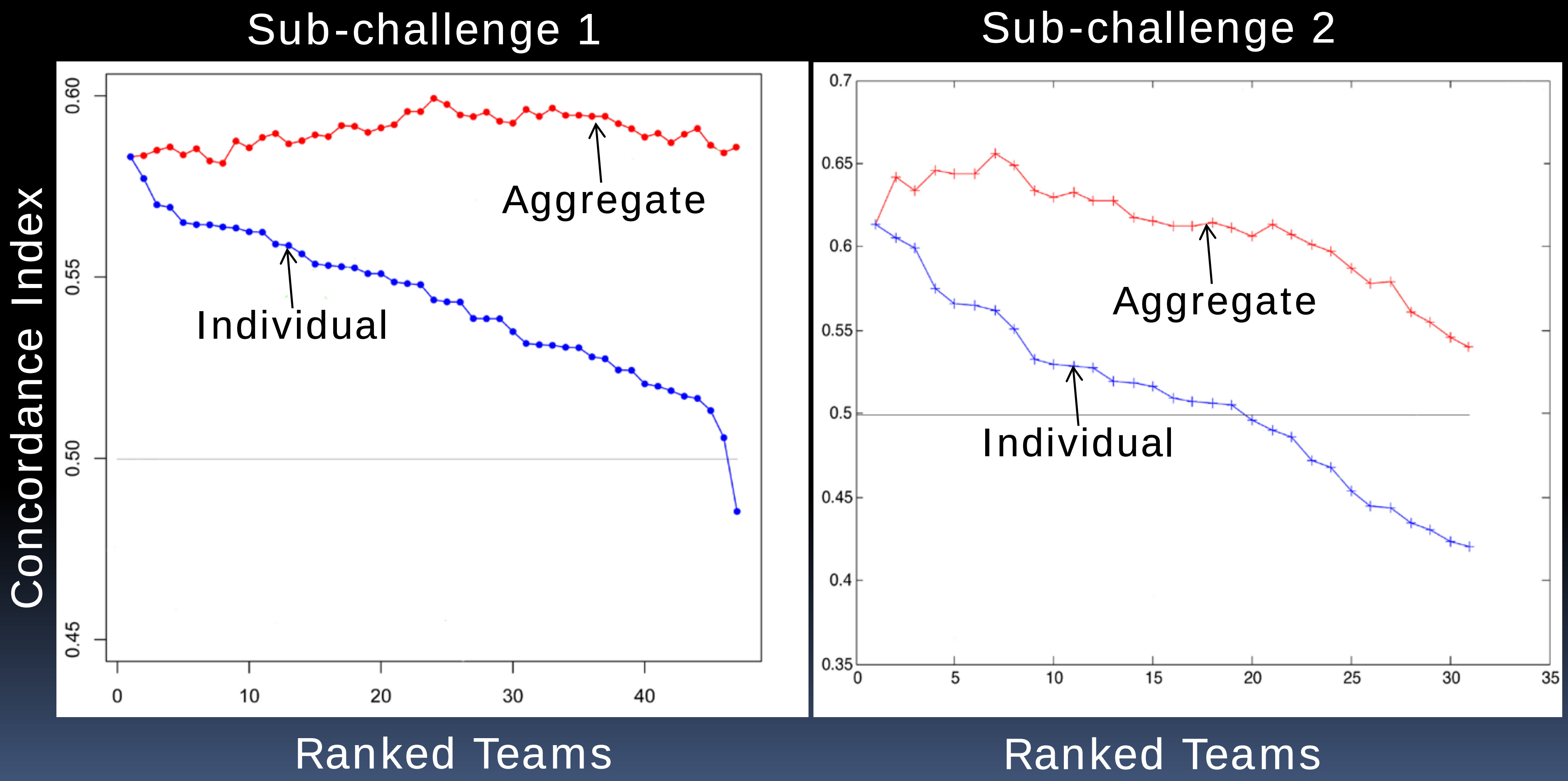
- Approach
  - Combining all data with additional prior knowledge
  - Gene set views
  - Discretized views, i.e., Binary conversion
  - Non-linear regression, multitask learning, Bayesian inference

Sub-Challenge 2:

UTSW-MC: University of Texas Southwestern Medical Center- Dallas,
TX, Jichen Yang and colleagues

- Approach
  - Combining all data with additional data sets
  - Matrix analysis of similarity between treatment “a” and “b”
  - Used only “growth” genes
  - Non-supervised approach
  - 8 pathways, 835 genes

Aggregation of results: The wisdom of the crowds



Sub-challenge 1

Sub-challenge 2

Next step for NCI-DREAM Challenge

- Further validation (Internal NCI- DREAM Team)
  - Sub-challenge 1: Additional breast cancer cell Lines from Joe Gray's lab
  - Sub-challenge 2: Test model on another lymphoma cell line
- Support winners to continue
  - Refining and enhancing their models, "hardening" and documenting software, making tools available to community
- Challenge assisted peer-review
  - Winners are writing an article about their winning model to Nat. Biotech, which was pre-approved to go to review

Lessons Learned

- Challenges:
  - Many approaches can be tested quickly and cheaply by clearly framing the problem and providing test and training data in well-defined format
- Community:
  - Hundreds to thousands of computationally sophisticated groups around the world will try to solve well-posed questions – even though some of them may miss the background to pose the questions themselves
  - Comparison of multiple approaches by crowdsourcing will accelerate learning in systems biomedicine and outcome optimization
- Models:
  - The wisdom of the crowd almost invariably outperformed that of individual teams
  - Not all computational approaches work equally well and we are still in early stages of identifying best approaches
  - Better performing approaches are those trained on other publically available data

Acknowledgements

- Sage Bionetworks
  - Stephen Friend
  - Adam Margolin
  - Erich Huang
  - Mike Kellen
  - Thea Norman
- Columbia University
  - Andrea Califano
  - Mukesh Bansal
  - Chuck Karan
- OHSU
  - Joe Gray
  - Laura Heiser
- NCI
  - Dinah Singer
  - Dan Gallahan
- DREAM
  - Gustavo Stolovitzky (IBM)
  - Erhan Bilal (IBM)
  - Jim Costello, BU
  - Julio Saez Rodriguez, EBI
  - Michael Menden, EBI
  - Thomas Cokelaer, EBI
- All DREAMers
  - From more than 40 different countries and 100 Institutions

Conclusions and Discussion

- What have we learned about data and models?
  - Challenges provide strong rationale for making well-curated data sets, computational platforms, and evaluation frameworks publically available
  - Wisdom of the crowd is a powerful mechanism to select tools of general value to the research community
  - Challenges help focus the attention of hundreds of researchers on relevant problems in need of analytical/computational solution
- Future challenges
  - To predict whether an *in vitro* study will or will not be validated in a pre-clinical context?
  - To predict *in vivo* compound toxicity? Efficacy? Outcome of clinical trials?
  - To predict genetic, transcriptional or metabolic interactions