What can integration of bioinformatic data tell us about disease relationships?

Erin Oerton[1,2], Ian Roberts[2], Tim Guilliams[2], Andreas Bender[1,2]

1. Centre for Molecular Informatics, Department of Chemistry, University of Cambridge
2. Healx Ltd, Mount Pleasant House, Mount Pleasant, Cambridge CB3 0RN

Background: the What, How, and Why of bioinformatic data integration

Relationships between diseases can be defined on multiple levels, from the observable phenotype down to molecular-level changes. Combining data across these diverse feature spaces may give greater insight into disease biology[1,2,3]. However, it is not yet clear:

- How biological data can best be integrated across such varied feature spaces
- Whether data integration will lead to more 'informative' disease connections than the use of different data types in isolation.

In this work we explored connectivity between 86 biologically diverse diseases, evaluating the predictive ability of individual feature spaces compared to kernels constructed from data fusion, as well as testing the impact of different linkage and kernel construction methods.

Using disease similarity to integrate biological data types

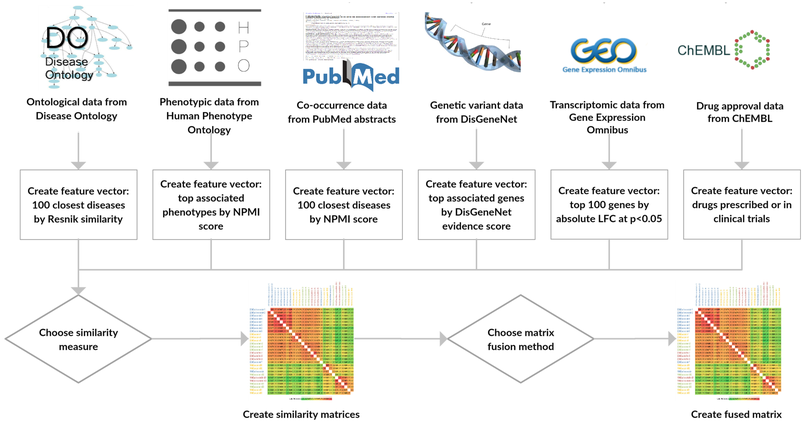


Fig. 1 Disease data for 86 diseases were obtained from six different sources, spanning different levels of systems biology. These data were then used to define similarities between diseases using one of three different similarity metrics, resulting in the construction of a similarity matrix. The matrices were then fused using one of seven fusion methods.

Differing predictive ability of data types and fusion methods

The 'discriminative power' (the difference between the similarity scores of related diseases and those of unrelated diseases) of data types varies greatly (error bars, Fig. 2). Phenotype- and literature co-occurrence–based kernels have the strongest discriminative ability for both DO and drug-sharing relationships. Drug-sharing relationships are a good predictor of ontological relationships, but the reverse is not true.

Discriminative power also varies for different similarity metrics and fusion methods. The results shown on this poster use Jaccard similarity with median average fusion, which demonstrated the greatest increase in predictive ability compared to other methods such as sum or rank fusion. A naïve Bayes classifier built using this kernel achieved 80% repeated random sub-sampling accuracy in a classification task against top-level DO classes.

Discriminative power increases with kernel size

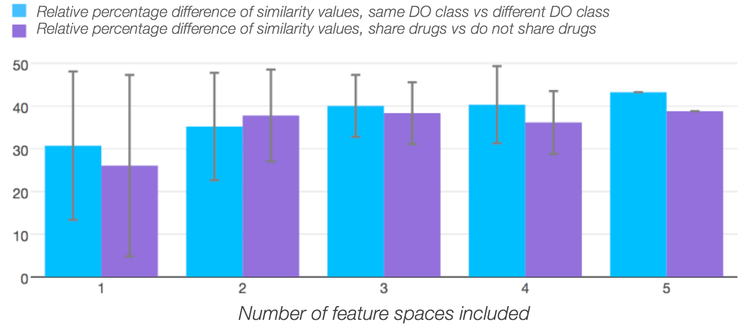
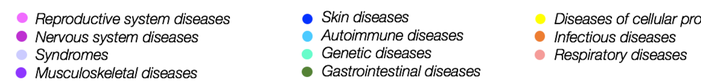


Fig. 2 The idea of combining feature spaces is appealing, as it allows multiple data types to be represented in a single kernel. Given the differing predictive abilities of each space, one concern is that this data fusion might result in a loss of predictive power. However, the predictive power of the fused kernel is actually greater than the average power of each space individually, showing that there is complementarity in the relationships in each feature space.

New disease relationships span ontological classes



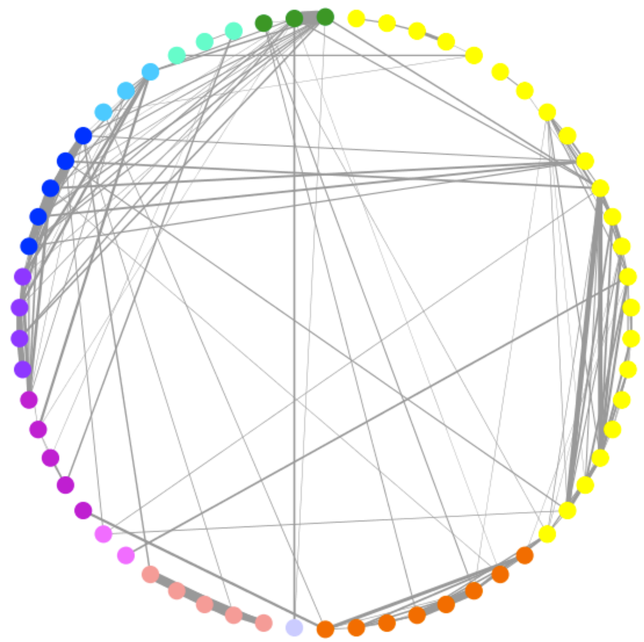


Fig. 3 Kernels can be used to define a 'disease map', where diseases are connected above a certain threshold of similarity. This figure shows the disease map constructed from all six feature spaces at a cutoff of 0.95, i.e. the top 5% of disease pairs by similarity value. The map defined by this kernel relates diseases not only within traditional disease classes (with strong interconnections in particular between cancers, infectious, and respiratory diseases) but also shows connections across ontological classes, such as between gastrointestinal and skin diseases.

Novel disease relationships are predictive of drug-sharing

Kernels constructed from fused data connect diseases across traditional ontological classes (as defined by DO). To evaluate the biological relevance of these novel relationships, drug-sharing was used as a proxy to understand whether there might be a shared biological mechanism underlying the connection. 32% of the novel disease connections in the full kernel have shared a drug in clinical trials, compared to an average of 24% of the novel connections in the individual kernels.

Case study: diseases related to psoriasis

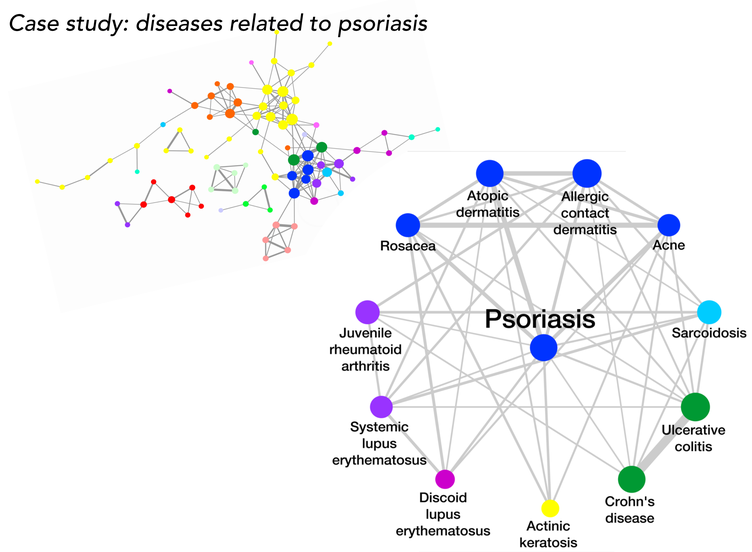


Fig. 4 The disease map defined by the fused kernel illustrates that the strongest relationships of psoriasis are to skin diseases, as expected, and other autoimmune diseases. Less expected are the links between psoriasis and the inflammatory bowel diseases. Examination of these links reveals shared dysregulated genes including the pro-inflammatory S100 family and CXC chemokines (associated with immune system activation). These diseases also share several drugs such as methotrexate, prednisone, and the psoriasis drug secukinumab, which has reached Phase II for Crohn's disease.

Fusing data results in greater predictive ability

Data integration not only is able to reconstruct known disease and drug-sharing relationships, but offers the intriguing possibility of highlighting novel relationships, which may eventually help to drive drug repurposing[2,3]. Future work will consider:

- Whether these conclusions apply equally across disease categories
- The impact of missing data (where data for a particular disease is available in certain spaces but not others) on the conclusions that can be drawn from integrated data.

References and Acknowledgements

Thanks to Dezső Módos for help with network visualization, and to the BBSRC Doctoral Training Programme in Bioscience for PhD funding.

1. Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N (2013) Discovering disease-disease associations by fusing systems-level molecular data. Sci Rep 3:3202
2. Sun K, Buchan N, Larminie C, Przulj N (2014) The integrated disease network. Integr Biol 6:1069–1079
3. Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, Sun F, Waterman M, Chen JJ, Chaudhary PM (2014) DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Res.