Semantic Biomarker Selection for Functional Genomics of Heart Failure Model Organisms

Ludwig Lausser[1], Steffen Just[2], Wolfgang Rottbauer[2], Hans A. Kestler[1,3]

[1] Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany
[2] Department of Internal Medicine II, Ulm University, Germany
[3] Leibniz Institute on Aging – Fritz Lipmann Institute, 07745 Jena, Germany

Abstract

Genetic model organism have the potential of increasing the understanding of malicious human genotypes and deregulated molecular processes. In this context, an adequate high-level characterization of experimental results is a necessary requirement for a successful trans-species transfer. In classification experiments, we analyze the gene expression profiles of six heart failure phenotypes of the zebrafish danio rerio. We train semantic multi classifier systems that directly provide a high-level hypotheses on the underlying processes.

1. Introduction

The molecular underpinnings of human heart failure are poorly defined, mainly due to significant mortality, low percentage of familial forms and limited access to cardiac tissue. Genetic model organisms like, mice and zebrafish can support the understanding of the genetic etiology of this disease. Selection of molecular markers and pathways is an essential step in the identification of possible disease causes at the molecular level.

We present a semantic multi-classifier system, which incorporates existing domain knowledge in the biomarker selection process. We construct interpretable marker subsets by using known relationships of measurements to higher-level terms such as pathways, e.g. P53-signalling. The subset is then used for training an expert (or base classifier). Vocabularies for these high-level terms are extracted from databases like KEGG [1] and Gene Ontology [2]. Our semantic multi-classifier system then selects the subset of terms with highest performance, see Figure 1.

2. Methods

Classification. A classifier is a predictive model that predicts the class label (e.g. phenotype) of an object $y \in \mathcal{Y}$ according to a set of measurements $\mathbf{x} \in \mathbb{R}^n$. It is typically adapted in a data-driven procedure and afterwards tested on an independent set of test samples. The training of a classifier system is typically coupled to an internal feature selection process in order to increase the interpretability of the final classification model [3]. It can be used for generating hypotheses on the underlying mechanisms leading to a specific phenotype.

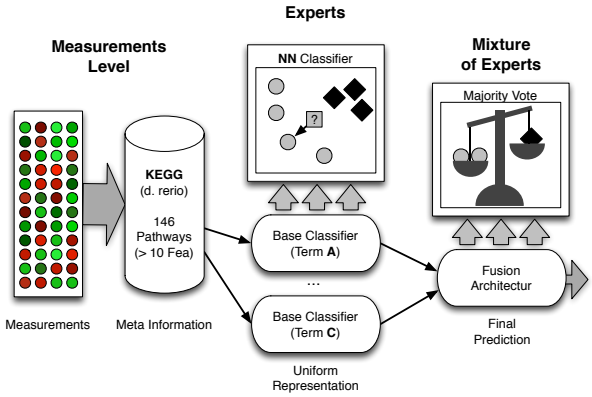

Figure 1. Classification and semantic integration scheme. Measurements and meta-information are integrated via a mixture of experts approach. Meta-information defines the features for the base classifiers and serves as a means of process identification. Final predictions are performed through majority votes.

In this work, we concentrate on the one nearest neighbor classifier (1-NN) [4]. The 1-NN identifies the training sample closest to the query sample (Euclidean distance). The label of this nearest neighbor is used as a prediction of the query sample's class label. More precisely, we apply a majority vote ensemble based on three 1-NN classifiers operating on different data representations, which will be selected in a semantic biomarker selection process [5]. An overview of the complete classification scheme is given in Figure 1.

 ISSN: 2325-887X DOI:10.22489/CinC.2017.258-195

Semantic Biomarker Selection. Semantic biomarker selection incorporates existing domain knowledge in the process of generating reduced marker sets [5, 6]. It assumes that the components of high-level processes such as signaling pathways are known a priori. The semantic biomarker selection process receives a vocabulary of high-level terms and restricts the feature space to those components. The prediction model can subsequently be directly interpreted via the terms selected. In our experiments, the three terms with the individual best accuracies (3×3 nested cross-validation [7]) were chosen for the ensemble classifier. As a vocabulary KEGG-pathways were utilized.

KEGG-Pathways. The Kyoto Encyclopedia of Genes and Genomes (KEGG) provides a catalog of known signaling pathway for different model organisms [1]. In case of danio rerio, it comprises 162 signaling pathways. Here, we restricted ourselves to those pathways for which our dataset provides more than 10 components (146 pathways).

2.1. RNA Sequencing Data

RNAs from 26 different heart failure zebrafish mutant lines identified in large-scale ENU-mutagenesis screens were isolated and subjected to RNA Sequencing. Each sample represents the pool of at least 25 zebrafish embryos. Embryos were phenotyped and collected between 48 and 120 hours post fertilization (hpf).

An overview on the sample collection can be found in Table 1. Each sample is represented by an RNA-Seq profile of 31953 features (genes). The dataset comprises 180 samples of 26 distinct genotypes. Overall 90 mutants and 90 controls are available. For each genotype at least 3 mutants and 3 controls were collected. For our analysis, the samples were regrouped into coarser (partially overlapping) phenotypical categories (mutant/controls), for each RNA sequencing was performed:

a) *bradycardia/arrythmia* (24/24): 7 different zebrafish mutant lines displaying severely reduced heart rates (bradycardia) and/or arrhythmia (atrial fibrillation, AV blocks).

b) *heart development* (66/66): 19 different zebrafish mutant lines displaying developmental heart defects such as defected heart chamber morphogenesis, cardiac differentiation, cardiac maturation.

c) *heart valve defect* (18/18): 6 different zebrafish mutant lines displaying defective heart valve formation.

d) *hypoplasia* (27/27): 8 different zebrafish mutant lines displaying reduced heart growth/reduced cardiomyocyte numbers due to diminished cardiomyocyte proliferation.

e) *myofibrillogenesis defects* (30/30): 7 different zebrafish mutant lines displaying defective myofibrillogenesis and myofibrillar organisation.

f) *weak contractility* (81/81): 23 different zebrafish mutant lines displaying severely reduced cardiac contractile functions.

Table 1. Overview on the analyzed dataset. The dataset consists of 180 samples of 26 individual genotypes (rows). The genotypes are named according to their observable heart beat. For each genotype the number of (mutants/controls) is reported. The genotypes were grouped into 6 overlapping phenotypes (columns).

samples(90/90)	bradycardia/ arrythmia (24/24)	developmental heart defects (66/66)	heart valve defect (18/18)	hypoplasia (27/27)	myofibrillogenesis defects (30/30)	weak contractility (81/81)
baldrian (3/3)	x	x	x	x		x
beach bum (3/3)		x	x			x
breakdance (3/3)	x					
bungee (3/3)		x	x	x		
dead beat (3/3)						x
flatline (3/3)		x			x	x
heart of stone (3/3)		x				x
herzbuckel (3/3)		x		x		x
island beat (6/6)	x	x		x		x
lazy susan (3/3)						x
liebeskummer (3/3)		x				x
lost contact (3/3)						x
main squeeze (3/3)						x
ping pong (3/3)		x	x	x		x
reggae (3/3)	x					x
schneckentempo (3/3)	x					x
schneeball (3/3)		x				x
steif (3/3)		x			x	x
silent heart (3/3)		x			x	x
tell tale heart (3/3)		x			x	x
titin (12/12)		x			x	x
titin-heart specific (3/3)		x			x	x
trapped (3/3)	x	x	x	x		x
weak atrium (3/3)		x			x	x
weiches herz (3/3)	x	x	x	x		x
windbeutel (3/3)		x		x		

3. Experimental Setup

The classifiers are tested in 10×10 cross-validation experiments [7]. That is the overall set of samples is split into ten folds of approximately equal size. Nine of these folds are used to train a classification model. The tenth fold is used as an independent test set. The procedure is repeated for each fold and the average accuracy, sensitivity, specificity are reported. Cross-validation is repeated for ten permutations of the set of samples. Experiments were performed with the TunePareto Software [8].

4. Results

The result of the 10×10 cross-validation experiments are given in Table 2. Accuracies, sensitivities (mutant)

and specificities (control) are shown. In general, the analyzed dichotomies (mutant/control) achieved high classification accuracies of 95.7% or higher. An exception is the *bradycardia/arrythmia* phenotype with a lower accuracy of 87.3%. Similar observations hold for sensitivities and specificities.

Table 2 additionally provides the three most frequently selected terms. For all phenotypes despite of the *bradycardia/arrythmia* phenotype, high-level terms could be identified that were selected in more than 93.0% of the corresponding experiments. For the general discrimination of mutant and control phenotypes the terms *fatty acid elongation* (93%), *butanoate metabolis* (92.0 %) and *biosynthesis of unsaturated fatty acids* (44.0 %) were selected most frequently. The term *fatty acid elongation* was passed to six individual phenotypes ($\geq$ 79%). *Butanoate metabolis* can be found in three other phenotypes ($\geq 55.0\%$) and *biosynthesis of unsaturated fatty acids* in four ($\geq 58.0\%$). The experiment with the individual phenotypes revealed five additional frequently selected terms: *beta-Alanine metabolism*, *dorso-ventral axis formation*, *insulin signaling pathway*, *melanogenesis*, *tyrosine metabolis*.

Table 2. Results of the 10×10 cross-validation experiments. The average accuracy, sensitivity (mutation) and specificity (control) is reported. Additionally the three most frequently selected KEGG-pathways are reported (%).

Mutation (90/90)		
Acc: 96.3%	Sens: 94.9%	Spec: 97.7%
1. Fatty acid elongation (93.0 %)		
2. Butanoate metabolism (92.0 %)		
3. Biosynthesis of unsaturated fatty acids (44.0 %)		
Bradycardia/Arrythmia (24/24)		
Acc: 87.3%	Sens: 82.5%	Spec: 92.1%
1. Tyrosine metabolism (60.0%)		
2. beta-Alanine metabolism (35.0%)		
3. Insulin signaling pathway (34.0%)		
Developmental heart defects (66/66)		
Acc: 98.7%	Sens: 98.3%	Spec: 99.1%
1. Fatty acid elongation (98.0 %)		
2. Biosynthesis of unsaturated fatty acids (93.0 %)		
3. Butanoate metabolism (55.0 %)		
Heart valve defect (18/18)		
Acc: 100.0%	Sens: 100.0%	Spec: 100.0%
1. Fatty acid elongation (100.0 %)		
2. Tyrosine metabolism (100.0 %)		
3. Dorso-ventral axis formation (22.0 %)		
Hypoplasia (27/27)		
Acc: 95.7%	Sens: 95.9%	Spec: 95.6%
1. Fatty acid elongation (100.0 %)		
2. Biosynthesis of unsaturated fatty acids (58.0 %)		
3. Melanogenesis (54.0 %)		
Myofibrillogenesis defects (30/30)		
Acc: 100.0%	Sens: 100.0%	Spec: 100.0%
1. Biosynthesis of unsaturated fatty acids (95.0 %)		
2. Fatty acid elongation (79.0 %)		
3. Butanoate metabolism (70.0 %)		
Weak contractility (81/81)		
Acc: 98.5%	Sens: 98.5%	Spec: 98.4%
1. Fatty acid elongation (99.0 %)		
2. Butanoate metabolism (97.0 %)		
3. Biosynthesis of unsaturated fatty acids (71.0 %)		

5. Discussion & Conclusion

In this work, we analyzed RNA-seq profiles of six different heart failure phenotypes gained of the model organism danio rerio. Our semantic multi classifier systems were able to achieve high accuracies ($\geq 87.3\%$) for all analyzed dichotomies. The performance of the classifier systems were coupled to stable term selections. Individual terms were selected in up to 100.0% of the experiments. In this way the semantic multi classifier systems generated new high-level hypotheses for the individual phenotypes, which could not be derived by purely data-driven analysis.

An detailed analysis of the found high-level terms and pathways is still ongoing work. The gene sets and the individual genes associated to these pathways are currently being investigated by additional wet lab experiments. Nevertheless, additional support and evidence can be found in the literature.

One of the most striking results is that the "Fatty Acid Elongation" KEGG-term seems to play a central role in discriminating mutation from not mutated specimens. This might be linked to $\omega - 3$ fatty acids. Also there exist recommendations from the American Heart Association on the consumption $\omega - 6$ polyunsaturated fatty acids due to its risks in cardiovascular diseases [9]. The benefit of $\omega - 3$ fatty acid supplementation is controversially discussed [10]. Several reviews on this topic exist [11, 12]. Jump et al. [12] report on the influence of polyunsaturated fatty acids on pathways that are involved in the regulation of blood lipids, inflammatory factors and the cellular processes in cardiomyocytes and vascular endothelial cells.

Riehle and Abel provide a review on insulin signaling and heart failure in homo sapiens [13]. In humans, heart failure is associated to insulin resistant states such as type 2 diabetes and obesity. In cardiomyocytes, changes in insulin signaling lead to the failing heart. Gao et al. showed that insulin sensitivity is increased in mice that receive a high-fat diet (butyrate) and counteracts insulin resistance phenotypes [14]. DeBosch and Muslin review the influence of insulin signaling pathways on cardiac growth [15].

The β alanine metabolism is known to play an important role in the training of the skeletal muscle [16]. Its influence on the heart muscle is investigated by dietary β alanine presupplementation in rats [17]. The respective animals showed a 57% reduction in infarct size to risk area ratio [18].

The semantic multi classifier model proposed in this study is prototypical. It can be extended in different ways. Especially alternative sources of domain knowledge can be utilized, which focus on different aspects of the phenotypes analyzed. Other specialized data bases might reveal additional processes that are not reflected in the KEGG database. A subsequent step is can be the transfer to different model organisms and may help to shed light on specific heart failure processes.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°602783, the German Research Foundation (DFG, SFB 1074 project Z1), the Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SySta R, ID 0315894A) all to HAK. SJ and HAK have received funding from the Federal Ministry of Education and Research (BMBF, e:Med, SYMBOL-HF, ID 01ZX1407A).

References

[1] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 2000;28(1):27–30.

[2] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G. Gene Ontology: tool for the unification of biology. Nature Genetics 2000; 25(1):25–29.

[3] Lausser L, Müssel C, Kestler HA. Measuring and visualizing the stability of biomarker selection techniques. Comput Stat 2013;28(1):51–65.

[4] Fix E, Hodges JL. Discriminatory analysis: Nonparametric discrimination: Consistency properties. Technical Report Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas, 1951.

[5] Lausser L, Schmid F, Platzer M, Sillanpää MJ, Kestler HA. Semantic multi-classifier systems for the analysis of gene expression profiles. Arch Data Sci Ser A Online First 2016; 1(1).

[6] Taudien S, Lausser L, Giamarellos-Bourboulis EJ, Sponholz C, F. S, Felder M, Schirra LR, Schmid F, Gogos C, S. G, Petersen BS, Franke A, Lieb W, Huse K, Zipfel PF, Kurzai O, Moepps B, Gierschik P, Bauer M, Scherag A, Kestler HA, Platzer M. Genetic factors of the disease course after sepsis: Rare deleterious variants are predictive. EBioMedicine 2016;12:227–238.

[7] Japkowicz N, Shah M. Evaluating Learning Algorithms: A Classification Perspective. New York: Cambridge University Press, 2011.

[8] Müssel C, Lausser L, Maucher M, Kestler HA. Multi-objective parameter selection for classifiers. J Stat Software 2012;46(5):1–27.

[9] Harris WS, Mozaffarian D, Rimm E, Kris-Etherton P, Rudel LL, Appel LJ, Engler MM, Engler MB, Sacks F. Omega-6 fatty acids and risk for cardiovascular disease. Circulation 2009;119(6):902–907.

[10] Mohebi-Nejad A, Bikdeli B. Omega-3 supplements and cardiovascular diseases. TANAFFOS 2014;13(1):6–14.

[11] Surette ME. The science behind dietary omega-3 fatty acids. Canadian Medical Association Journal 2008; 178(2):177–180.

[12] Jump DB, Depner CM, Tripathy S. Omega-3 fatty acid supplementation and cardiovascular disease. thematic review series: New lipid and lipoprotein targets for the treatment of cardiometabolic diseases. Journal of Lipid Research 2012; 53(12):2525–2545.

[13] Riehle C, Abel ED. Insulin signaling and heart failure. Circulation Research 2016;118(7):1151–1169.

[14] Gao Z, Yin J, Zhang J, Ward RE, Martin RJ, Lefevre M, Cefalu WT, Ye J. Butyrate improves insulin sensitivity and increases energy expenditure in mice. Diabetes 2009; 58(7):1509–1517.

[15] DeBosch BJ, Muslin AJ. Insulin signaling pathways and cardiac growth. Journal of Molecular and Cellular Cardiology 2008;44(5):855 – 864.

[16] Sale C, Saunders B, Harris RC. Effect of beta-alanine supplementation on muscle carnosine concentrations and exercise performance. Amino Acids 2010;39(2):321–333.

[17] McCarty MF, DiNicolantonio JJ. β-alanine and orotate as supplements for cardiac protection. Open Heart 2014;1(1).

[18] Allo SN, Bagby L, Schaffer SW. Taurine depletion, a novel mechanism for cardioprotection from regional ischemia. American Journal of Physiology Heart and Circulatory Physiology 1997;273(4):H1956–H1961.

Address for correspondence:

Hans A. Kestler
Institute of Medical Systems Biology, Ulm University
89069 Ulm, Germany
hans.kestler@uni-ulm.de / hans.kestler@leibniz-fli.de