Revealing comprehensive genotype–phenotype associations through logic relationships

Alexey V. Antonov, Hans W. Mewes

GSF National Research Center for Environment and Health
Institute for Bioinformatics

Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

antonov@gsf.de
w.mewes@gsf.de

Abstract: A novel approach which employs principles of higher order logic analyses was developed to systematically correlate phylogenetic data with phenotype profiles by identification of phenotype specific patterns of presence of multiple proteins. For example, for most genomes expressing trait A, the presence of protein C presumes the presence of protein B, while for other genomes (not expressing the trait) the presence of protein C presumes the absence of protein B. We demonstrate that the phenotype specific patterns reflect fundamental structural changes in the genotype of microorganisms in relation to conditions provided by presence/absence of a trait. We discover many previously unidentified genotype–phenotype associations on the level of fundamental biochemical processes.

1 Introduction

Microbial species express a variety of phenotypic traits and behaviors. Previous studies have shown that the molecular basis of these traits can be partially understood by application of comparative genomics. It was demonstrated that genes whose phylogenetic profiles correlate well with a given microbial phenotypic profile often belong to the same biochemical pathway directly related to the trait [1-6]. However, such simple approach does not take into account the complexity of cellular networks [7, 8] and, thus, in most cases do only partially reflect genome variations between phenotypes analyzed.

To extend and generalize this approach we propose to combine principles of higher order logic analyses with phenotype data to get a deeper understanding of genotype-phenotype associations. In simple terms, in addition to identification of single genes we are also looking for gene pairs and triplets that are present only in species with a given phenotype while in other species the genes are rarely present together. We refer to such protein pairs and triplets as complex (phenotype) patterns.

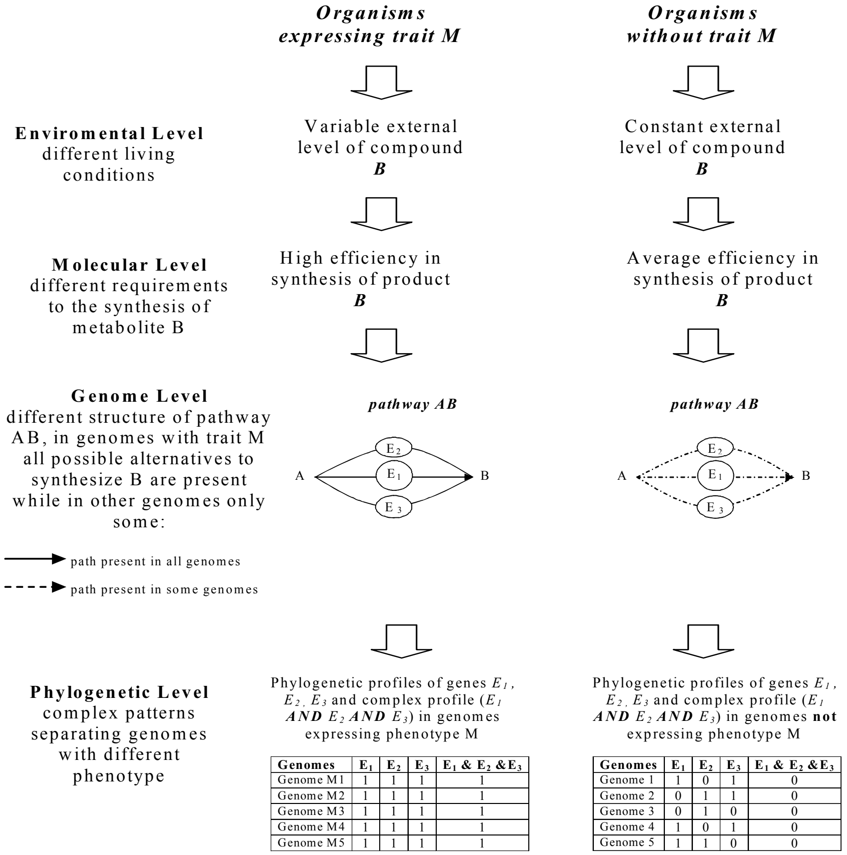
What is the logic for the existence of complex phenotype patterns? The logic for high order relationships inside a group of genes conserved in the genomes expressing a specific phenotype and absent in other genomes is similar to the originally proposed scientific foundation described by Bowers et al. [7, 8]. In our case, it is related to phenotype specific pathways or phenotype specific parts of the pathways. As we demonstrate further, in most cases these genome modifications are not detectable in the data based on pairwise similarity of genes phylogenetic and phenotypic profiles. Consideration of higher order logic involving complex gene patterns is required and returns highly valuable information.

Our results support that genes involved in complex phenotype patterns are functionally linked (from the same or closed biochemical pathway(s)). In many cases the identified biochemical pathways were related to fundamental processes suggesting profound genome reorganization between analyzed phenotypes.

2 The biological basis of complex phenotype patterns

The species that show variation in the expression of a trait were demonstrated to have phenotype specific pathways. The profiles of these genes correlate significantly to the phenotype profile and can be easily recovered by pairwise similarity analyses. This fact was supported by a number of studies [1, 4]. We refer to such genome variations as primary, with respect to their visibility on the level of single gene profiles. However, it is natural to expect extra genome variations of pathways related to fundamental processes like metabolism, energy, transport. These variations are triggered by different environmental conditions which are imposed on species from different phenotypes. On the molecular level, this is reflected by different requirements to the efficiency of the same biochemical processes. On the genome level, it is likely to expect the different structural organization of the corresponding pathways (figure 1).

In most cases evolution implements a number of different alternatives to realize biochemical processes. The most preferable alternative for a particular organism that would be selected under evolutionary pressure depends on many factors (environmental conditions, availability in the genome of other biochemical pathways and so on). For example, for a group of species expressing a particular trait, it may be vital to have all alternatives to synthesize some metabolite, while for others the availability of one single alternative is sufficient for survival. As a result, we expect proteins catalyzing the synthesis to be joined in phenotype specific patterns. Such structural reorganization of genome content between species of different phenotypes is visible only on the level of complex phenotype patterns in most cases.


Organisms expressing trait M
Organisms without trait M
Enviromental Level
different living conditions
Variable external level of compound B
Constant external level of compound B
Molecular Level
different requirements to the synthesis of metabolite B
High efficiency in synthesis of product B
Average efficiency in synthesis of product B
Genome Level
different structure of pathway AB, in genomes with trait M all possible alternatives to synthesize B are present while in other genomes only some:
pathway AB
pathway AB
A
E2
E1
E3
B
A
E2
E1
E3
B
path present in all genomes
path present in some genomes
Phylogenetic Level
complex patterns separating genomes with different phenotype
Phylogenetic profiles of genes E1, E2, E3 and complex profile (E1 AND E2 AND E3) in genomes expressing phenotype M
Phylogenetic profiles of genes E1, E2, E3 and complex profile (E1 AND E2 AND E3) in genomes not expressing phenotype M

Genomes	E_1	E_2	E_3	E_1 & E_2 &E_3
Genome M1	1	1	1	1
Genome M2	1	1	1	1
Genome M3	1	1	1	1
Genome M4	1	1	1	1
Genome M5	1	1	1	1

Genomes	E_1	E_2	E_3	E_1 & E_2 &E_3
Genome 1	1	0	1	0
Genome 2	0	1	1	0
Genome 3	0	1	0	0
Genome 4	1	0	1	0
Genome 5	1	1	0	0

Figure 1: An illustrative example for emergence of complex phenotype patterns. The presence/absence of trait M has caused different evolutionary pressure on the pathway AB due to the different requirements to the efficiency of synthesis of product B. As a result, genomes that express phenotype M in most cases have all possible pathways AB while genomes that do not express phenotype M have only one (maybe two) pathways. The signature of such structural genome variations between phenotypes is a presence of (E_1 **AND** E_2 **AND** E_3) pattern. The profiles of genes E_1 , E_2 , E_3 is not (in general) deferentially present between phenotypes. Therefore, differences of genome content between phenotypes is visible only on the level of complex patterns.

Consider a simple illustrative example such as a pathway of transforming metabolite A into metabolite B. Let us assume that only three different enzymes (E1, E2, E3) exist to transform metabolite A into metabolite B (figure 1) and the species expressing trait M due to environmental conditions utilize product B more intensively. Let us also assume that the synthesis of product B is more efficient in the case if all three enzymes (E_1, E_2, E_3) are present in the organism. Under these assumptions the species expressing trait M and having all alternative enzymes will get preferences for selection. Therefore, in most species expressing trait M all alternative pathways AB are most probably available while most organisms without trait M have only some alternative pathways. If we consider the phylogenetic profiles of enzymes (E_1, E_2, E_3) that catalyze pathway AB then we observe the "AND" phenotype pattern (E_1 & E_2 & E_3). At the same time the single profile of each enzyme (E_1, E_2, E_3) is not necessarily differentially present between phenotypes (see (figure 1).

3 Methods

The phylogenetic profile of an arbitrary gene A is formalized as a binary vector g (elements of g can be either 1 or 0, indicating whether the homolog of gene A is present in the corresponding genome or not). The phenotype information is similarly formalized as a binary vector f (elements of f can be either 1 or 0, indicating whether the corresponding genome expresses the considered phenotype or not). We refer to each gene profile g as a base (single) profile to differentiate between single gene profiles and complex profiles (see below). We also refer to base profiles as complex profiles of the first degree.

Consider a vector $C(g_A , g_B) = (g_A \cap g_B)$ which represents a binary profile indicating presence of both gene A and gene B in a genome. We refer to such vectors as complex profiles. In general we can use three logical operations ("AND", "EXCLUDE", "OR"). Here we consider only "AND" logical operation. Each complex phylogenetic profile is characterized by the number of base profiles required to construct it. We refer to this characteristic as the degree. For example, we already defined base profiles as complex profiles of the first degree. The complex profile $g_A \cap g_B$ is a complex profile of the second degree (gene pair), the profile $C(g_A , g_B, g_C) = ((g_A \cap g_B) \cap g_C)$ is a third degree profile (gene triplet).

The relation between a phenotype f and a genotype g_A is quantified by a similarity measure I. The empirical mutual information was proved to be the best choice [1, 3]. An arbitrary complex profile $C(g_A , g_B, g_C)$ can be related to the phenotype f in the same way by similarity measure I.

3.1 Extraction of complex patterns

As input the set of phylogenetic profiles related to the genome of interest and the profile f related to the analyzed phenotype is employed. Single profiles g_i whose similarity I_i to the phenotype profile f is greater than a given threshold I_1 are selected. These genes are expected to represent primary genetic variations related to the phenotype. We refer to this set as GF_1 (index 1 indicates the degree of phylogenetic profiles). These profiles are dropped from consideration while searching for the higher degree (complex) profiles.

In the next step we checked all possible combinations of remaining gene pairs to identify complex profile of the second degree whose similarity to phenotype profile exceeds threshold value I_2. We refer to this set as GF_2 (index 2 indicates the degree of phylogenetic profiles). These profiles are also dropped from consideration while searching for the set of third degree profiles.

To identify complex patterns of the third degree (gene triplets that differentially present between organisms of different phenotype f) we applied our searching algorithm[9]. Those third degree patterns whose similarity to phenotype f was found to be significant (was better then threshold value I_3) were selected to the set GF_3 (index 3 indicates the degree of complex profiles).

At each step we dropped from consideration significant low level complex profiles. The reason for this is the same as been argued earlier[7, 8]. For example, for each triplet of genes a, b, c that was suspected to be a pattern they require that neither profile a nor b alone was predictive of c.

The threshold values I_1, I_2, I_3 are identified based on the background distribution of similarity measure I for patterns of the first, second and third degree respectively. The background distribution was computed based on random simulation procedure. The random simulation procedure was repeated 500 times. In the first step each time (k) the random phenotype f_k was generated by random permutation of the bits (genomes) in the phenotype vector f. In the next step the set of phylogenetic profiles and the random phenotype f_k is used as input to infer complex patterns related to random phenotype f_k (for procedure see above). The best profiles (in respect to similarity measure I) of the first, second and third degree that were identified are selected and their similarity measures I_{1k}, I_{2k} , I_{3k} are accumulated. The distributions I_{1k} , I_{2k}, I_{3k} , $k = 1..500$ are used as background distributions of similarity measure I for profiles of the first, second and third degree respectively.

3.2 Analyses of biological relevance of complex phenotype patterns

All unique genes that were linked in significant patterns (sets GF_2, GF_3) were selected into the separate gene cluster (genes that were involved in less then 5 patterns were dropped from consideration). In the next step the cluster was analyzed by automatic functional profiling. Automatic functional profiling is the standard procedure for the analysis of biological relevance of a gene set [9-11] or gene networks[12, 13]. Given a set A (the set of genes identified to be related to some biological phenomena) and a set B (reference set, usually the set of all genes from the analyzed organism), automatic functional profiling identifies statistically over/under represented attributes f (f is usually a functional category from employed annotation vocabulary F, i.e. GO[14] , FunCat[15]. If attribute f is over/under represented in set A it is said to be enriched. The knowledge of enriched functional categories is helpful for the understanding of the biological model and mechanism that unite the genes from the set A.

The identified gene clusters were profiled using FunCat[15] annotation. The genes from the cluster were considered as set A. The set of all genes from the analyzed genome was considered as the reference set B. To account for multiple testing the statistical significance of the enrichment was computed by the Monte-Carlo simulation approach (to adjust p-value for multiple testing). The estimated p-value corresponds exactly to the definition of an experiment–wise Westfall and Young p-value [16].

4 Results

As the framework for our analysis, we use two well studied bacterial genomes: E.coli and B.subtilis. We used these genomes as a benchmark to check reliably whether or not complex phenotype patterns are formed by functionally related genes. Functional catalog at MIPS [17] contains approximately 80% of manually annotated genes in E.coli genome and 75% genes from B.subtilis genome. The manual comprehensive genome annotation guarantees consistent functional analysis of genes from identified patterns. For all genes from E.coli and B.subtilis genomes the gene phylogenetic profiles in approximately 200 complete microbial genomes were acquired. The profiles were downloaded from http://tavazoielab.princeton.edu/genphen/.

We investigated the genetic basis of morphological traits, namely genome variations that took place between gram-negative and gram-positive bacteria. E.coli is a gram-negative bacterium while B.subtilis is a gram-positive one. Gram-negative bacteria are in general characterized by the presence of an additional membrane layer, the OM that serves as a permeability barrier to prevent the entry of toxic compounds while allowing the influx of nutrient molecules [18].

In the first step we identified 202 genes in E.coli genome and 262 genes in B.subtilis genome whose single gene phylogenetic profiles correlate significantly (p-value < 0.01) to the gram-negative and gram-positive phenotype respectively. Functional profiling of these genes revealed significant (p-value < 0.01) enrichment of functional categories. Most of categories related to gram-negative case are directly linked to the analyzed phenotype (table 1). Categories like "bacterial outer membrane only in Gram- bacteria" reflect on the genomic level the differences between two phenotypes: gram-negatives develop additional membrane layer. These genotype–phenotype associations were identified previously [1].

	Category Code	**Category Name**	**Set A statistics**		**Set B statistics**		**P-value***	**P-value corrected for multiple testing**
1	42.34.01	bacterial outer membrane only in Gram- bacteria	*11*	*202*	*9*	*4981*	2.77E-11	< 0.01

2	42	BIOGENESIS OF CELLULAR COMPONENTS	*25*	*202*	*179*	*4981*	8.54E-08	< 0.01
3	32.05	disease, virulence and defense	*19*	*202*	*110*	*4981*	2.34E-07	< 0.01
4	32	CELL RESCUE, DEFENSE AND VIRULENCE	*32*	*202*	*305*	*4981*	2.86E-07	< 0.01
5	14	PROTEIN FATE folding, modification, destination	*34*	*202*	*347*	*4981*	4.19E-07	< 0.01
6	01.20.15.03	biosynthesis of ubiquinone	*5*	*202*	*1*	*4981*	4.92E-07	< 0.01
7	14.04	protein targeting, sorting and translocation	*10*	*202*	*26*	*4981*	5.80E-07	< 0.01
8	70.34.01	bacterial outer membrane only present in Gram- bacteria	*12*	*202*	*47*	*4981*	1.52E-06	< 0.01
9	32.05.01	resistance proteins	*15*	*202*	*85*	*4981*	3.74E-06	< 0.01
10	01.20.15	biosynthesis of derivatives of dehydroquinic acid, shikimic acid and chorismic acid	*5*	*202*	*4*	*4981*	9.17E-06	< 0.01

Table 1. Functional terms enriched (p-value < 0.01) in the set of 202 genes from *E.coli* genome. Single phylogenetic profiles of these genes are significantly associated with gram-negative phenotype. Functional terms in the table are ordered by statistical significance. Columns "Set A statistics" and "Set B statistics" disclose detailed statistics for each enriched category in the selected gene set (202 genes) and in the remaining genes from the whole genome (4981 genes).

* Column indicates p-value (hypogeometric test) without correction for multiple testing. To account for multiple testing the statistical significance was computed by the Monte-Carlo simulation approach.

The enriched functional terms related to gram-positives are not phenotype specific (table 2). This means that gram-positives develop no specific (in relation to gram-negatives) biochemical processes but significantly modify the existing ones by gaining new genes. For example, more then 30 percent (13 out 36) of genes related to *"phosphotransferase system"* in *B.subtilis* genome are specific for gram-positive genomes only.

	Category Code	**Category Name**	**Set A statistics**		**Set B statistics**		**P-value***	**P-value corrected for multiple testing**
1	16.03	nucleic acid binding	*39*	*262*	*187*	*3782*	9.95E-10	< 0.01
2	16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT structural or catalytic	*68*	*262*	*556*	*3782*	8.60E-09	< 0.01
3	20.03.09	phosphotransferase system	*13*	*262*	*23*	*3782*	1.28E-07	< 0.01
4	16.03.01	DNA binding	*30*	*262*	*160*	*3782*	8.07E-07	< 0.01

5	01.06.10	regulation of lipid, fatty acid and isoprenoid metabolism	*5*	*262*	*0*	*3782*	1.09E-06	< 0.01
6	01.05.04	regulation of C-compound and carbohydrate utilization	*11*	*262*	*27*	*3782*	1.34E-05	< 0.01

Table 2. Functional terms enriched (p-value < 0.01) in the set of 262 genes from *B.subtilis* genome. Single phylogenetic profiles of these genes are significantly associated with gram-positive phenotype. Functional terms in the table are ordered by statistical significance. Columns “Set A statistics” and “Set B statistics” disclose detailed statistics for each enriched category in the selected gene set (262 genes) and in the remaining genes from the whole genome (3782 genes).

* Column indicates p-value (hypogeometric test) without correction for multiple testing. To account for multiple testing the statistical significance was computed by the Monte-Carlo simulation approach.

Application of our computational approach to E.coli genome detects gene pairs and triplets that present together only in gram-negative bacteria while analysis of B.subtilis genome reveals gene pairs and triplets that present together only in gram-positive genomes. Thus, by analyses of these patterns we track variations in the structure of biochemical processes between gram-negative and gram-positive genomes. These variations are not triggered by gain of new genes but by different joint distribution between phenotypes of several phenotype unspecific genes.

We identified 2714 significant patterns (p-value < 0.01) of the second degree (gene pairs) and 1756 significant patterns (p-value < 0.01) of the third degree (gene triplets) specific for gram-negative and 6234 gene pairs and 860 gene triplets specific for gram-positive. We considered only “AND” patterns (pairs and triplets of genes joined by “AND” logical operation) and at each step removed from further consideration significant patterns identified at previous steps (while looking for gene pairs we dropped from consideration 202 genes (gram-negative) and 262 genes (gram-positive) which represent patterns of the first degree, while looking for gene triplets we dropped from consideration identified patterns of the second degree (see methods)). The sets of significant patterns (pairs and triplets) for each case were transformed into the clusters GC_{neg} and GC_{pos} which consist of all unique genes that were linked in significant patterns identified for gram-negative and gram-positive phenotype respectively. We dropped from consideration genes that take part in less then 5 patterns. The size of the clusters was equal to 299 and 405 genes respectively.

In both cases, the genes from the identified clusters reflect variations in the structure of biochemical processes t between gram-negative and gram-positive genomes. Indeed, we found that the clusters were enriched (p-value < 0.01) by several functional terms. Tables 3 and 4 supplies detailed information for all functional categories enriched in each cluster.

	Category Code	Category Name	Set A statistics		Set B statistics		P-value*	P-value corrected for multiple testing
1	16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT structural or catalytic	*127*	*299*	*987*	*4884*	5.67E-25	< 0.01
2	01	METABOLISM	*118*	*299*	*1148*	*4884*	2.44E-16	< 0.01
3	70.03	cytoplasm	*66*	*299*	*426*	*4884*	1.61E-13	< 0.01
4	14	PROTEIN FATE folding, modification, destination	*48*	*299*	*333*	*4884*	8.04E-09	< 0.01
5	01.01.06.06.01	biosynthesis of lysine	*7*	*299*	*3*	*4884*	1.98E-07	< 0.01
6	02.11	electron transport and membrane-associated energy conservation	*21*	*299*	*93*	*4884*	8.26E-07	< 0.01
7	16.03.03	RNA binding	*19*	*299*	*78*	*4884*	1.11E-06	< 0.01
8	02.11.05	accessory proteins of electron transport and membrane-associated energy conservation	*12*	*299*	*28*	*4884*	1.14E-06	< 0.01
9	01.01.06.06	metabolism of lysine	*7*	*299*	*5*	*4884*	1.16E-06	< 0.01
10	01.01.06	metabolism of the aspartate family	*11*	*299*	*26*	*4884*	3.55E-06	< 0.01
11	11.04	RNA processing	*10*	*299*	*21*	*4884*	4.44E-06	< 0.01
12	42.01	cell wall	*10*	*299*	*22*	*4884*	6.09E-06	< 0.01
13	01.03.16	polynucleotide degradation	*8*	*299*	*14*	*4884*	1.55E-05	< 0.01
14	01.03.16.01	RNA degradation	*6*	*299*	*6*	*4884*	2.26E-05	< 0.01

Table 3. Functional terms enriched (p-value < 0.01) in the cluster GC_{neg} (299 unique genes from *E.coli* genome that were linked in significant patterns identified for gram-negative genomes). Functional terms in the table are ordered by statistical significance. Columns “Set A statistics” and “Set B statistics” disclose detailed statistics for each enriched category in the selected gene set (299 genes) and in the remaining genes from the whole genome (4884 genes).

* Column indicates p-value (hypogeometric test) without correction for multiple testing. To account for multiple testing the statistical significance was computed by the Monte-Carlo simulation approach.

Among 299 genes that were involved in complex patterns for gram-negative phenotype we found significant enrichments related to several general functional categories, like, “metabolism” or “PROTEIN WITH BINDING FUNCTION”. Some specific functional categories, like, “electron transport and membrane-associated energy conservation” were enriched as well. Our results suggest that in addition to biochemical processes that were significantly modified in gram-negative bacteria by gain of new genes (table 1) there was additional significant reorganization of the structure of some processes reflected by categories in table 3.

	Category Code	Category Name	Set A statistics		Set B statistics		P-value*	P-value corrected for multiple testing
1	20.09.18	cellular import	*82*	*405*	*161*	*3639*	9.71E-28	< 0.01
2	70.30	prokaryotic cytoplasmic membrane	*114*	*405*	*403*	*3639*	5.18E-23	< 0.01
3	20.09	transport routes	*93*	*405*	*266*	*3639*	9.31E-23	< 0.01
4	30	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	*52*	*405*	*74*	*3639*	1.08E-21	< 0.01
5	16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT structural or catalytic	*122*	*405*	*502*	*3639*	8.92E-21	< 0.01
6	01	METABOLISM	*157*	*405*	*842*	*3639*	1.37E-20	< 0.01
7	30.01	intracellular signalling	*44*	*405*	*57*	*3639*	1.77E-19	< 0.01
8	30.01.01	unspecified signal transduction	*41*	*405*	*48*	*3639*	2.89E-19	< 0.01
9	20.01	transported compounds substrates	*86*	*405*	*306*	*3639*	3.02E-16	< 0.01
10	30.01.05	enzyme mediated signal transduction	*31*	*405*	*31*	*3639*	5.24E-16	< 0.01
11	20.03.25	ABC transporters	*54*	*405*	*136*	*3639*	1.94E-14	< 0.01
12	70.03	cytoplasm	*103*	*405*	*472*	*3639*	2.80E-14	< 0.01
13	20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	*98*	*405*	*431*	*3639*	2.93E-14	< 0.01
14	30.05.01.10	two-component signal transduction system sensor kinase component	*21*	*405*	*16*	*3639*	1.39E-12	< 0.01
15	20.03	transport facilitation	*67*	*405*	*237*	*3639*	1.99E-12	< 0.01
16	14.07.03	modification by phosphorylation, dephosphorylation, autophosphorylation	*32*	*405*	*55*	*3639*	5.44E-12	< 0.01
17	20.01.07	amino acid transport	*27*	*405*	*38*	*3639*	1.08E-11	< 0.01
18	14.07	protein modification	*47*	*405*	*137*	*3639*	6.98E-11	< 0.01
19	30.05	transmembrane signal transduction	*21*	*405*	*28*	*3639*	1.20E-09	< 0.01
20	14	PROTEIN FATE folding, modification, destination	*57*	*405*	*221*	*3639*	2.39E-09	< 0.01
21	20.01.03	C-compound and carbohydrate transport	*32*	*405*	*81*	*3639*	8.71E-09	< 0.01
22	11.02.03.04	transcriptional control	*58*	*405*	*240*	*3639*	1.07E-08	< 0.01
23	11.02	RNA synthesis	*58*	*405*	*245*	*3639*	1.92E-08	< 0.01
24	11.02.03	mRNA synthesis	*58*	*405*	*245*	*3639*	1.92E-08	< 0.01
25	01.03	nucleotide metabolism	*32*	*405*	*85*	*3639*	2.12E-08	< 0.01

26	16.21	complex cofactor/cosubstrate binding	*28*	*405*	*68*	*3639*	4.08E-08	< 0.01
27	16.21.17	pyridoxal phosphate binding	*13*	*405*	*13*	*3639*	2.18E-07	< 0.01
28	16.11	amino acid binding	*7*	*405*	*1*	*3639*	6.86E-07	< 0.01
29	16.17	metal binding	*23*	*405*	*57*	*3639*	9.24E-07	< 0.01
30	01.03.01.03	purine nucleotide anabolism	*9*	*405*	*5*	*3639*	1.08E-06	< 0.01
31	20.01.01.01.01.01	siderophore-iron transport	*8*	*405*	*4*	*3639*	3.04E-06	< 0.01
32	01.01.03.02.01	biosynthesis of glutamate	*7*	*405*	*5*	*3639*	4.46E-05	< 0.01

Table 4. Functional terms enriched (p-value < 0.01) in the cluster GC_{pos} (405 unique genes from *B.subtilis* genome that were linked in significant patterns identified for gram-positive genomes). Functional terms in the table are ordered by statistical significance. Columns “Set A statistics” and “Set B statistics” disclose detailed statistics for each enriched category in the selected gene set (405 genes) and in the remaining genes from the whole genome (3639 genes).

* Column indicates p-value (hypogeometric test) without correction for multiple testing. To account for multiple testing the statistical significance was computed by the Monte-Carlo simulation approach.

Among 405 genes that were involved in complex phenotype patterns for gram-positive phenotype we found significant enrichments related to several general functional categories, like, “CELLULAR COMMUNICATION”, “CELLULAR TRANSPORT”, “PROTEIN FATE”, “METABOLISM”. Our results suggest that transport processes were significantly modified in gram-positive genomes in comparison to gram-negative ones. Both clusters (GC_{pos} and GC_{neg}) were significantly enriched by proteins with binding function. These findings imply that different environment conditions between gram-positives and gram-negative phenotypes require different binding spectra of the cell proteome.

5 Discussion

We have extended the ideas underlying the logical analysis of phylogenetic profiles to the investigation of genomic and phenotype data. Previously the phylogenetic data were related to phenotype only by pairwise similarity (between phylogenetic and phenotype profiles). This undemanding approach does not take into account the complexity of cellular networks (branching, parallel, and alternate pathways). In most cases the simple correlation of genotype to phenotype identifies only primary genome variations which were directly caused by (or cause by themselves) the phenotype divergence. However, these changes reflect only a small fraction of reorganization of genome content that took place between phenotypes. Our approach makes visible additional genome variations that took place on the level of fundamental biochemical pathways and processes. Finally, we would like to point out that the proposed statistical structures (the patterns of multiple presence of proteins associated with phenotype) have not been explored until now. They are different from any previously explored statistical patterns in phylogenetic data.

References

1. Slonim N, Elemento O, Tavazoie S. Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Bio*; **2**:2006.0005.

2. Huynen M, Dandekar T, Bork P. Differential genome analysis applied to the species-specific features of Helicobacter pylori. *FEBS Lett* 1998; **426(1)**:1-5.

3. Jim K, Parmar K, Singh M, Tavazoie S. A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res* 2004; **14(1)**:109-115.

4. Levesque M, Shasha D, Kim W, Surette MG, Benfey PN. Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Curr Biol* 2003; **13(2)**:129-133.

5. Liu Y, Li J, Sam L, Goh CS, Gerstein M, Lussier YA. An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS Comput Biol* 2006; **2(11)**:e159.

6. Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD*, et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 2005; **3(5)**:e134.

7. Bowers PM, Cokus SJ, Eisenberg D, Yeates TO. Use of logic relationships to decipher protein network organization. *Science* 2004; **306(5705)**:2246-2249.

8. Bowers PM, O'Connor BD, Cokus SJ, Sprinzak E, Yeates TO, Eisenberg D. Utilizing logical relationships in genomic data to decipher cellular processes. *FEBS J* 2005; **272(20)**:5110-5118.

9. Antonov AV, Mewes HW. Complex functionality of gene groups identified from high-throughput data. *J Mol Biol* 2006; **363(1)**:289-296.

10. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005; **21(18)**:3587-3595.

11. Khatri P, Bhavsar P, Bawa G, Draghici S. Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res* 2004; **32(Web Server issue)**:W449-W456.

12. Antonov AV, Mewes HW. BIOREL: the benchmark resource to estimate the relevance of the gene networks. *FEBS Lett* 2006; **580(3)**:844-848.

13. Antonov AV, Tetko IV, Mewes HW. A systematic approach to infer biological relevance and biases of gene network structures. *Nucleic Acids Res* 2006; **34(1)**:e6.

14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM*, et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25(1)**:25-29.

15. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G*, et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* 2004; **32(Database issue)**:D41-D44.

16. Westfall PH, Zaykin DV, Young SS. Multiple tests for genetic effects in association studies. *Methods Mol Biol* 2002; **184**:143-168.

17. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M*, et al.* The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 2004; **32(18)**:5539-5545.

18. Nikaido H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* 2003; **67(4)**:593-656.