BMC Bioinformatics

RESEARCH ARTICLE Open Access

Conserved genomic neighborhood is a strong but no perfect indicator for a direct interaction of microbial gene products

Robert Esch[1] and Rainer Merkl[2*]

Abstract

Background: The order of genes in bacterial genomes is not random; for example, the products of genes belonging to an operon work together in the same pathway. The cotranslational assembly of protein complexes is deemed to conserve genomic neighborhoods even stronger than a common function. This is why a conserved genomic neighborhood can be utilized to predict, whether gene products form protein complexes.

Results: We were interested to assess the performance of a neighborhood-based classifier that analyzes a large number of genomes. Thus, we determined for the genes encoding the subunits of 494 experimentally verified hetero-dimers their local genomic context. In order to generate phylogenetically comprehensive genomic neighborhoods, we utilized the tools offered by the Enzyme Function Initiative. For each subunit, a sequence similarity network was generated and the corresponding genome neighborhood network was analyzed to deduce the most frequent gene product. This was predicted as interaction partner, if its abundance exceeded a threshold, which was the frequency giving rise to the maximal Matthews correlation coefficient. For the threshold of 16%, the true positive rate was 45%, the false positive rate 0.06%, and the precision 55%. For approximately 20% of the subunits, the interaction partner was not found in a neighborhood of ± 10 genes.

Conclusions: Our phylogenetically comprehensive analysis confirmed that complex formation is a strong evolutionary factor that conserves genome neighborhoods. On the other hand, for 55% of the cases analyzed here, classification failed. Either, the interaction partner was not present in a ± 10 gene window or was not the most frequent gene product.

Keywords: Protein-protein interaction, Complex formation, Sequence similarity network, Genome neighborhood network, Binary classifier

Background

A fundamental organizational unit of microbial genomes is the operon consisting of a cluster of genes that are transcribed into a single mRNA molecule [1, 2], which allows for the quasi-parallel synthesis of the gene products. Commonly, proteins encoded in the same operon work together, e. g., as enzymes catalyzing subsequent steps of a metabolic pathway. Thus, genomic neighborhood is a reliable indicator for the functional association of proteins and an important element for the generation of functional networks offered by databases like STRING [3].

An early comparison of nine bacterial and archeal genomes has led to the conclusion that proteins encoded by conserved gene pairs interact physically [4]. An example is the *trp* operon of *Escherichia coli* that consists of the five genes *trp*A – *trp*E catalyzing tryptophan biosynthesis from chorismate [5]. In *E. coli*, *trp*C encodes as a fusion of two genes a bifunctional protein that has TrpC and TrpF functionality. As Fig. 1a shows, the five *trp* genes occur in the same order also in the closely related γ-Proteobacterium *Salmonella enterica* enterica A3ES40 and in the genome of *Bacillus subtilis* subtilis 168, where TrpC and TrpF are encoded by two separate genes. In stark contrast, the BioCyc database [6] indicates that *trp*E lies isolated from other *trp* genes in the genome of *Agrobacterium* sp. H13–3 and that the distance

* Correspondence: rainer.merkl@ur.de
[2] Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany
Full list of author information is available at the end of the article

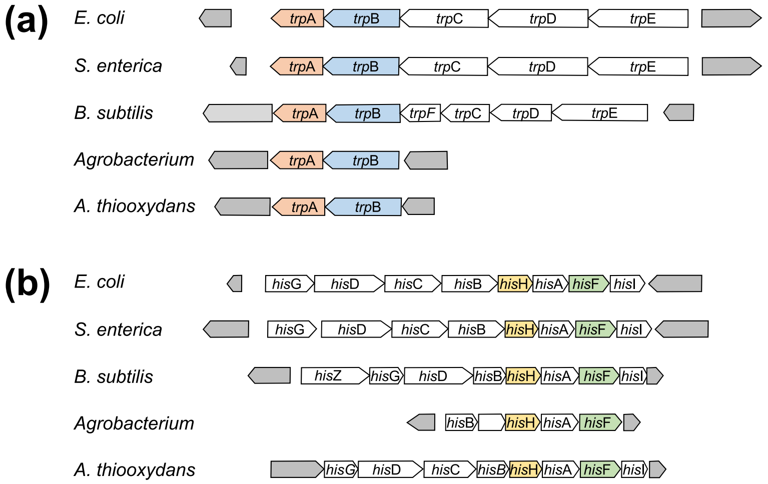

Fig. 1 Genomic neighborhood of the *trp*A gene (**a**) and of the *his*F gene (**b**) in five bacterial genomes. The neighborhoods were deduced from the BioCyc database [6] for the genomes of *E. coli* K-12, *S. enterica* enterica A3ES40, *B. subtilis* subtilis 168, *Agrobacterium* sp. H13–3, and *A. thiooxydans* ZJ. The output of BioCyc is shown schematically, but drawn to scale; neighboring genes not related to tryptophan or histidine biosynthesis are filled grey. The gene products of *trp*A and *trp*B and of *his*H and *his*F (all color-coded) form hetero-oligomers, respectively. The conservation of the *trp*A and *trp*B neighborhood in all five genomes suggests the formation of a TrpA/TrpB complex. For the same genomes, the neighborhood of the *his*H gene would more likely propose a HisH/HisA/HisF, a HisH/HisA, or a HisA/HisF, but no HisH/HisF complex. Note that HisH and HisF form a hetero-dimer, whereas HisA is a monomeric protein [7]

to *trp*A is > 2,200,000 bp. In the genome of *Acidiferrobacter thiooxydans* ZJ, the genomic neighborhood of *trp*E contains *pab*A (a *trp*G homolog), *trp*D, and *trp*C, but the distance to *trp*A is > 50,000 bp. However, in all five genomes, *trp*A and *trp*B are genomic neighbors and the proteins TrpA and TrpB are the two subunits that form the tryptophan synthase, which is a permanent, hetero-oligomeric protein complex that experienced an intricate evolutionary history [8]. These observations propose that a direct protein-protein interaction is an evolutionary factor that preserves genomic neighborhood considerably stronger than a functional interaction of the gene products. Indeed, cotranslational protein assembly and the order, in which the gene products assemble to a complex affect the order of genes in operons [9]; the latter effect is stronger for weakly expressed genes [10].

On the other hand, the comparison of *his* operons that consist of the genes leading to the synthesis of histidine [11] indicates some intricacies that may complicate a too simplistic inference of direct protein interactions. The *his* operon of *E. coli* contains eight genes and only the *his*H and the *his*F gene products form a hetero-dimer, whereas all other gene products are monomers. As shown in Fig. 1b, the comparison of the *his* operons from the five species introduced above, makes clear that their *his*H and *his*F genes are no immediate neighbors, but are separated by *his*A, which encodes a monomeric enzyme [7]. If this neighborhood of the three genes *his*H, *his*A, and *his*F is conserved in many genomes, it is not possible to deduce in silico the formation of the HisH/HisF complex.

In literature, we did not find a comprehensive characterization of a classifier that predicts the subunits of microbial protein complexes by analyzing genomic neighborhoods. Due to the more than 200,000 sequencing projects listed in the GOLD database [12], we expected a statistically comprehensive sampling of neighborhoods that is sufficient to determine the reliability of such a classifier. We concentrated on the assessment of experimentally verified hetero-dimers based on the analysis of large sets of genomes and addressed two specific questions:

1) How often are microbial proteins that form hetero-dimeric complexes encoded in close genomic vicinity?
2) How reliable is the prediction of hetero-dimeric protein complexes based on neighborhood conservation deduced from a comprehensive number of genomes?

Thus, we determined for the subunits of 494 hetero-dimers the genomic neighborhoods from phylogenetically diverse species and assessed the frequency of the most abundant gene products in a ± 10 gene window. It turned out that 80% of the known interaction partners are encoded in a ± 10 neighborhood. Additionally, we predicted for each subunit the product of the most frequent neighbor as direct interaction partner. Applying a threshold that balances false positive and false negative predictions, 485 of the 1087 known interaction partners were correctly identified by choosing the most abundant gene neighbor.

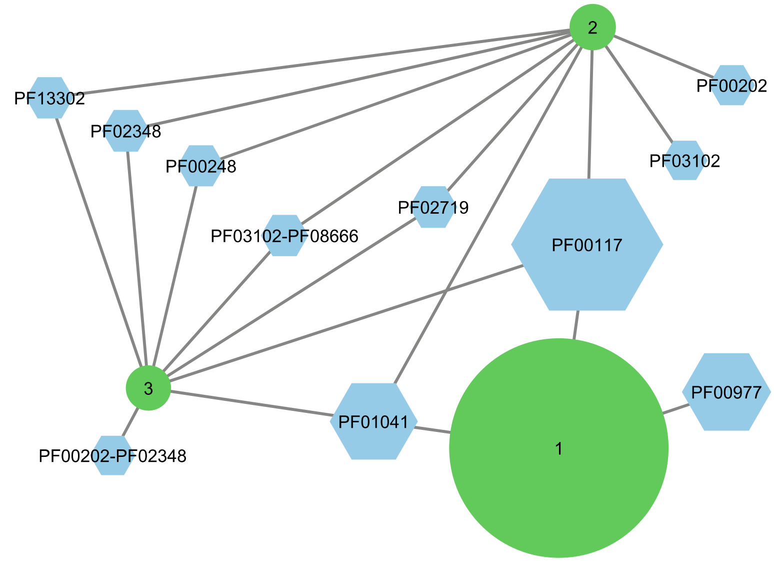

Fig. 2 rGNN of InterPro family IPR004651 characterizing the histidine biosynthesis protein HisF. The circles represent three *rep_nodes* 40; their diameter corresponds to the number of corresponding sequences. The hexagons represent *pf_nodes* (i. e., Pfam families) that are labeled with their Pfam-IDs; the size of the hexagons corresponds to the value of *f*(*pf_node*). These *pf_nodes* indicate the most abundant protein functions encoded in the ± 10 genomic neighborhood of the HisF sequences. PF00117 (glutamine amidotransferase I, HisH) was correctly identified as interaction partner of HisF, because it had the largest value $f^*_{\max}(pf\_node)$, which was 43.6%. To reduce complexity, all nodes representing ≤ 3 sequences were eliminated for this illustration

Results

Deducing phylogenetically diverse genomic neighborhoods of bona fide bacterial hetero-dimers

We were interested to find out how reliable the genomic neighborhood indicates for a given subunit (*su*) of a heteromeric complex the interaction partner, if we analyze its neighborhood in genomes from many phylogenetically diverse species. This approach relies on two, mutually reinforcing effects: *i*) Prokaryotic genomes are rather unstable [13, 14], only 5–25% of the genes belonging to operons are shared by at least two distantly related species [15]. Thus, in the framework of this analysis, the abundances of those genes must be low that encode in the neighborhood of a given *su* such proteins that are only functionally related or even functionally unrelated to the *su* under study. *ii*) In contrast, if the propensity for a direct protein-protein interaction affects genomic distances, the specific interaction partners must often be neighbors in a large number of phylogenetically diverse genomes and thus stand out through higher abundances.

The prediction method considered here can only be applied to hetero-oligomers. The simplest form of heteromeric complexes are hetero-dimers consisting of two subunits. Thus, we chose a recently compiled set of bacterial hetero-dimers with known crystal structures that do not possess additional interaction partners like DNA [16]. The corresponding PDB [17] entries were analyzed to deduce pairs of complex-forming subunits (su_i^1, su_i^2) and the corresponding InterPro [18] and Pfam [19] families. These annotations were indispensable for the subsequent analysis (see below); after the elimination of ambiguous cases, 494 pairs of subunits remained. The corresponding PDB-IDs and detailed results are listed in Additional file 1: Table S1.

To create for all su_i^* proteins neighborhoods that are phylogenetically most comprehensive, we utilized tools offered by the Enzyme Function Initiative (EFI), which were developed to analyze sequence and function space of protein families [20]. The EFI-Genome Neighborhood Tool (EFI-GNT) computes a genome neighborhood network (GNN) for a given sequence similarity network (SSN). This SSN has to be created beforehand by means of the EFI-Enzyme Similarity Tool (EFI-EST). Thus, for each of the su_i^* under study, we generated an SSN for the InterPro family it belonged to. If delivered by EFI-EST, we processed *rep_node* 80 networks, otherwise *rep_node* 40 files (for details see Methods).

We chose a ± 10 neighborhood (for justification see Methods) and created for each SSN, i. e. *rep_node* file, a refined genome neighborhood network (rGNN) by utilizing a modified version of AGeNNT [21]. In an rGNN, the neighboring gene products are represented by the Pfam families (*pf_nodes*) they belong to and the *pf_node*-specific *SeqCount* values indicate the number of neighborhoods encoding this protein function. The *SeqCount* values were transferred to relative frequencies *f*(*pf_* *nodes*); see Formula (1). For each su_i^*, we identified the *pf_node* with the highest frequency $f^*_{\max}(pf\_node)$, which was assumed to represent the putative interaction partner of su_i^*, if this frequency exceeded a lower threshold. We analyzed only *pf_nodes* occurring with a minimal frequency of 20%; thus, this analysis of rGNNs considered the large number of neighborhoods represented in the InterPro and Pfam databases, but additionally focused to the most frequent protein functions; see Fig. 2 for an example. Interestingly, the median *SeqCount* value for all su_i^* elements was 133 contributed by 21.5 phylogenetic phyla (median), which testifies for our analysis to a phylogenetically diverse representation of neighborhoods and genomes.

Assessing the abundance of neighboring interaction partners

Because some of the interacting proteins are composed of more than one domain, they belong to more than one InterPro family. Thus, our PDB-ID ↔ InterPro ↔ Pfam mapping resulted in 1087 subunits su_i^*, for which we knew the interaction partner involved in complex formation. These were the positive cases (P). Moreover, we assumed that these su_i^* do not form a complex with any of the other proteins encoded in the genomic neighborhood of su_i^*. These proteins summed up to 596,767 negative cases (N) and we used the P and N cases to compute performance values.

There are hetero-dimeric complexes, where the two subunits are not encoded in close genomic vicinity as exemplified by the *B. subtilis* enzymes PabA and TrpE that form the anthranilate synthase [16]. As indicated by BioCyc [6], *pab*A, which is a multipartner enzyme, is part of the pabBAC-sul-folBK-yazB-yacF-lysS transcription unit that starts in the *B. subtilis* genome at base position 82,831, whereas *trp*E belongs to the *trp* operon that begins at base position 2,377,619. Thus, we presumed that the number of false positive assignments decreases, if we introduce a minimal threshold frequency f_{min} by testing $f^*_{\max}(pf\_node) \geq f_{\min}$; see Formula (2). If a *pf_node* reaching f_{min} represented a known interaction partner, it was a true positive (TP), otherwise it was a false positive (FP). All other *pf_nodes* not reaching this threshold were false negative (FN), if representing an interaction partner and true negative (TN), if representing one of the negative cases.

The Matthews correlation coefficient (MCC) [22] is considered a fair performance measure even for unbalanced datasets, as it is deduced from all classified cases. Thus, we incremented the threshold f_{min} in 1% steps between 1 and 100% and determined f_{min}-specific MCC values; see Formula (4). As Fig. 3a indicates, the maximal MCC-value of 0.50 was achieved for f_{min} = 16%. In this case, 485 of the 1087 known interaction partners were predicted as TP and 389 of the 596,767 putatively non-interacting proteins were FP.

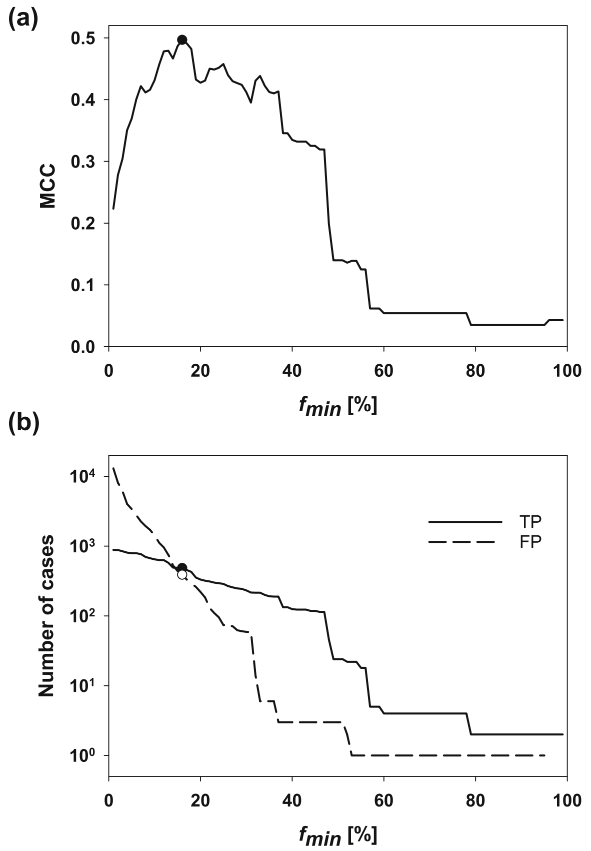

Fig. 3 MCC values versus the minimal threshold frequency f_{min} used to predict interaction partners (**a**). The f_{min} value (16%) resulting in the highest MCC is indicated by a filled circle. Absolute numbers of TP and FP predictions versus the minimal threshold frequency f_{min} (**b**). The two circles indicate the number of cases related to the chosen f_{min} value

Figure 3b confirms that the number of FPs is – as expected – negatively correlated with f_{min}, whereas the number of TPs is less affected by the chosen f_{min} value. Decreasing f_{min} below 16% would result in more FPs than TPs, which is not desired. Figure 4a represents the initial interval of a TPR versus FPR plot (ROC curve) ending at an FPR of 2%. This rate corresponds to 13,024 FPs, which is far above an acceptable performance. Choosing f_{min} = 16% results in a TPR of 45% and an FPR of 0.06% (see Formulae (3)); the latter rate must be chosen low due to the enormous number of 596,767 negative cases. Interestingly, the results indicate for this classifier a maximally TPR value of ~80%. This finding suggests for approximately 20% of the hetero-dimers that the subunits are not encoded in a ± 10 neighborhood. The Precision/Recall curve shown in Fig. 4b

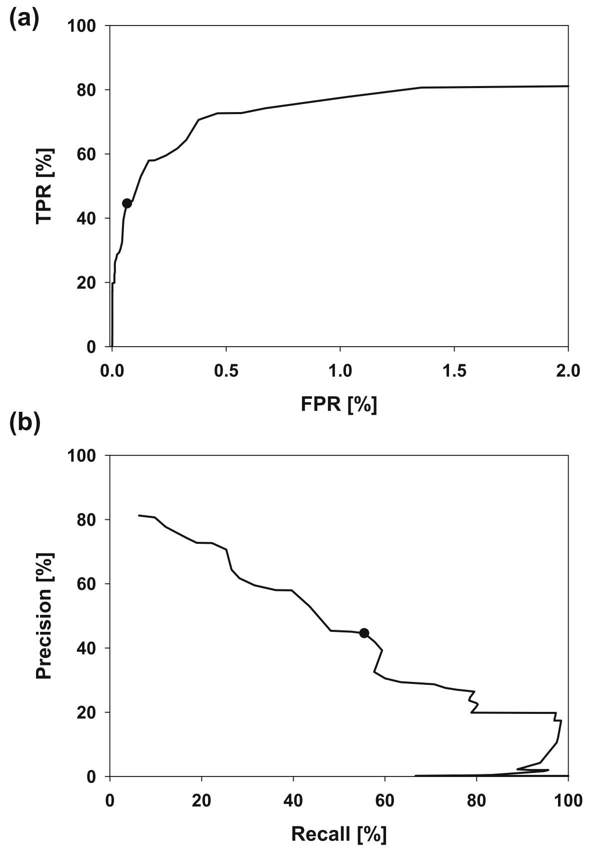

Fig. 4 TPR = Recall, FPR, and Precision for f_{min}-values incremented from 0 to 100%. Initial interval of a ROC curve ending at an FPR of 2% (**a**). The Precision/Recall plot for the full range of threshold values (**b**). TPR = Recall (45%), FPR (0.06%), and Precision (55%) values for the classifier operated with f_{min} = 16% are indicated by filled circles

indicates a nearly linear anticorrelation between Precision and Recall (i. e., the TPR). For the f_{min} value suggested by the MCC analysis, Precision is 55% and Recall 45%.

One might argue that smaller neighborhoods reduce the risk of predicting the wrong interaction partner. To address this problem, we have additionally analyzed the ± 3 neighborhood of those 27 cases, where the true interaction partner reached rank two in our prior classification. The results are summarized in Additional file 1: Table S2. For twelve cases (44%), the rank was unaltered; for seven cases (26%), the rank deteriorated below two and for eight cases (30%), the rank increased to one. This finding suggested to us that smaller neighborhoods have no pronounced effect on classification performance.

If several complexes are encoded in close vicinity, it is difficult to predict interaction partners

Our protocol identified 485 of the 1087 known interaction partners correctly, and we wanted to elucidate reasons for the prediction of 389 FP cases, which are *pf_nodes* (proteins) occurring in the GNNs with a higher frequency than the true interaction partner.

Among the false predictions was the interaction partner of SoxA, which forms together with SoxX the hetero-dimeric SaxAX cytochrome [23]. The gene of SoxY is located in the neighborhood of SaxA at varying positions, but the corresponding *pf_node* had rank three and the neighbors SoxY and SoxZ had higher frequencies, which led to the FP prediction of a SoxA:SoxZ interaction. SoxY and SoxZ form a complex together with SoxB and *sox*VWXYZABCDEFGH is a transcriptional unit in α-Proteobacteria [24]. The STRING database (version 11.0) indicated that the neighborhood of *soxX*, *soxY*, *soxZ*, and *soxA* is conserved in all α- and β-Proteobacteria that possess clear homologs of SoxA. This example illustrates that it is hard to predict the correct interaction partner, if the subunits of more than one complex are encoded in close vicinity.

Discussion

Limitations of the current approach

Although the median of the phyla contributing to the *SeqCount* values of the $f^*_{\max}(pf\_node)$ nodes was 21.5, one might argue that our approach overestimated the frequencies of functionally related or even functionally unrelated neighbors. The rigorous elimination of genomes from closely related species might reduce this bias, but not the one caused by the horizontal transfer of larger fragments like selfish operons [25]. More efficient would be an elimination method based on the pairwise comparison of the protein functions [26] encoded in the considered neighborhoods. Identical protein functions encoded in highly similar local arrangements would indicate closely related species or cases of horizontal gene transfer. However, this approach would require a rigid preprocessing of the genomes and a completely different software pipeline.

We used a dataset of 1087 subunits to determine the optimal MCC value and identified a threshold f_{min} of 16% as optimal. If one considers our algorithm as a classifier, one might argue that the algorithm's parameter were optimized and tested on the same dataset. A cross-validation technique could be used to exclude overfitting. However, as we fixed only one parameter (the f_{min} value), we consider the risk of overfitting minimal. The dataset analyzed here consists of proteins devoid of non-protein macromolecules that formed complexes with stoichiometries of AB, A_2B_2, A_3B_3, A_4B_4, A_6B_6, ABC, and $A_2B_2C_2$ [16]. As we analyzed only the full set of these proteins, the determined performance values might be misleading for test cases outside our training sample.

Conclusions

By using EFI services that process large datasets, we have confirmed for hetero-dimers that approximately 45% of the subunits are the most frequent gene products in the GNNs that correspond to a ± 10 neighborhood. Additionally, our data suggest that approximately 20% of the interaction partners are encoded outside of this genomic window.

A survey of the oligomerization state of *E. coli* proteins revealed that hetero-oligomers are a minority: 20% of the proteins are monomers, whereas dimers and tetramers are far more common; 79% of the complexes are homo-oligomers with 2 to 12 subunits and only 21% are hetero-oligomers [27]. Thus, for a comprehensive in silico prediction of all types of protein complexes, a machine learning approach combining several features is required in order to increase classification reliability; for a recent review see e. g. [28].

Methods

Mapping PDB entries to InterPro and Pfam families

For the mapping of chains from PDB datasets, the services offered by the European Bioinformatics Institute (EMBL-EBI) were used [29]. The pages "https://www.ebi.ac.uk/pdbe/entry/pdb/*$ID*/analysis" were parsed to determine for the proteins of the PDB dataset with PDB-ID *$ID* the InterPro and Pfam families. The PDB-IDs were taken from a recently prepared dataset [16] consisting of bona fide bacterial hetero-dimers. Mostly due to the co-existence of more than one domain, 186 subunits were mapped to more than one InterPro family, which were all analyzed. Only in one of these cases (chain A of the methylmalonyl-coa mutase PDB-ID 4req) the prediction varied among the assigned InterPro families; compare ranks in Additional file 1: Table S1.

Creating SSNs and GNNs

Our software pipeline consisted of scripts written in Python [30] that were executed on the compute-server of the EFI or an in-house computer, which were all equipped with Linux. All scripts are deposited at Github.

SSNs were computed command-line based on the EFI cluster for InterPro families with default parameters chosen by EFI-EST. An SSN consists of nodes each representing a sequence; the nodes are interconnected by edges weighted with the BLAST bit score resulting from a pairwise alignment of the related sequences. For large protein families, an extremely high number of edges renders an SSN intractable; thus, EFI-EST does not generate an output file, if the SSN would contain more than 10,000,000 edges. To reduce network complexity, EFI-EST maps sequences sharing at least *x*% sequence identity to one node and generates representative node (*rep_node*) *x* networks.

The SSNs were uploaded to the EFI server for the generation of GNNs. "Raw" GNNs were converted to rGNNs by means of an updated version of AGeNNT [21] that was adapted to the current EFI interfaces. We oriented ourselves on the architecture of the *E. coli* genome and a systematic analysis performed during the design of AGeNNT [21] to determine the window size of the genomic neighborhood to be analyzed. Approximately 80% of all *E. coli* transcription units have fewer than five genes and 80% of all directons, i. e., genes transcribed in the same direction with no intervening one transcribed in the opposite direction, have fewer than ten genes [31]. Moreover, a systematic screening indicated that ± 10 neighborhoods are best suited to identify gene clusters [21]. Thus, we selected a ± 10 neighborhood as default. In this case, the neighborhood consists for each member of the InterPro family of exactly those 20 gene products that are encoded in a ± 10 window. The output of the EFI-GNT is independent of the localization of transcriptional units and represents the function of these proteins by means of PFAM accession numbers.

Consequently, an rGNN consists of *rep_nodes*, i. e., a cluster of sequences from the InterPro family under study and *pf_nodes* representing enzyme functions encoded in the respective neighborhoods. For edges between *rep_nodes* and *pf_nodes*, the coverage is given, which is the relative number of neighborhoods containing the considered enzyme function represented by *pf_node*. For each *pf_node*, the *SeqCount* parameter indicates the number of genomic neighborhoods possessing this protein function. rGNNs are encoded as xgmml files, which were parsed to deduce from the *SeqCount* values the *pf_node* with the highest frequency $f^*_{\max}(pf\_node)$; this one was the candidate for the prediction of interaction partners.

Classifying *pf_nodes*

For the determination of a relative frequency *f(pf_node)*, the *pf_node*-specific *SeqCount(pf_node)* value was divided by the sum of all *SeqCount(pf_node*)* values observed in the rGNN under study, according to

$$f(pf\_node) = SeqCount(pf\_node) / \sum SeqCount(pf\_node*) \tag{1}$$

For the assignment of interaction partners, the frequency of the chosen *pf_node* was compared to the threshold f_{min} according to

$$pred(pf\_node) = \begin{cases} 1 & if\ f^*_{\max}(pf\_node) \geq f_{\min} \\ 0 & otherwise \end{cases} \tag{2}$$

and *pf_node* was considered as interacting, if *pred*(*pf_node*) was 1.

Performance measures

To assess the performance of a classification, we determined the false positive rate (FPR), the true positive rate (TPR = Recall), the Precision

$$FPR = \frac{FP}{N}, \quad Recall = TPR = \frac{TP}{P}, \quad Precision = \frac{TP}{TP + FP} \tag{3}$$

and the MCC value

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}. \tag{4}$$

In all formulae, P is the number of positive and N the number of negative cases. TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

Visualizing GNNs and rSSNs

All networks were visualized by means of Cytoscape [32].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3200-z.

Additional file 1. All raw data of the SSN/GNN analyses.

Abbreviations

EFI: enzyme function initiative; EFI-EST: EFI-enzyme similarity tool; EFI-GNT: EFI-genome neighborhood tool; FN: false negative cases; FP: false positive cases; FPR: false positive rate; MCC: Matthews correlation coefficient; N: negative cases; P: positive cases; PDB-ID: protein database identifier; pf_node: a node representing a Pfam family; Pfam-ID: Pfam database identifier; rGNN: refined genome neighborhood network; SeqCount: number of genomic neighborhoods encoding a certain function; SSN: sequence similarity network; su: subunit of a protein complex; TN: true negative cases; TP: true positive cases; TPR: true positive rate

Acknowledgements

We thank John A. Gerlt (University of Illinois) and his team for excellent support.

Authors' contributions

RM designed the project. RE developed the algorithms, ran the computational analysis, and prepared the output. RM wrote the manuscript, which was finalized with input from RE. All authors read and approved the final manuscript.

Funding

No funding was received for this study.

Availability of data and materials

The code of the software pipeline and the scripts for statistical analysis can be found at https://github.com/merkllab/ConsGNN

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Mathematics and Computer Science, University of Hagen, D-58084 Hagen, Germany. [2]Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040 Regensburg, Germany.

Received: 5 July 2019 Accepted: 8 November 2019
Published online: 03 January 2020

References

1. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3:318–56.
2. Rocha EP. The organization of the bacterial genome. Annu Rev Genet. 2008;42:211–33.
3. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003;31:258–61.
4. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 1998;23:324–8.
5. Yanofsky C. The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis*. Trends Genet. 2004;20:367–74.
6. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2014;42:D459–71.
7. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science. 2000;289:1546–50.
8. Merkl R. Modelling the evolution of the archeal tryptophan synthase. BMC Evol Biol. 2007;7:59.
9. Natan E, Wells JN, Teichmann SA, Marsh JA. Regulation, evolution and consequences of cotranslational protein complex assembly. Curr Opin Struct Biol. 2017;42:90–7.
10. Wells JN, Bergendahl LT, Marsh JA. Operon gene order is optimized for ordered protein complex assembly. Cell Rep. 2016;14:679–85.
11. Alifano P, Fani R, Lio P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev. 1996;60:44–69.
12. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A, Chen IA, Kyrpides NC, Reddy T. Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res. 2019;47:D649–D59.
13. Mushegian AR, Koonin EV. Gene order is not conserved in bacterial evolution. Trends Genet. 1996;12:289–90.
14. Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. 2009;6:23–52.
15. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 2001;11:356–72.
16. Plach MG, Semmelmann F, Busch F, Busch M, Heizinger L, Wysocki VH, Merkl R, Sterner R. Evolutionary diversification of protein-protein interactions by interface add-ons. Proc Natl Acad Sci U S A. 2017;114:E8333–E42.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.
18. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47:D351–D60.
19. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–D32.
20. Gerlt JA. Tools and strategies for discovering novel enzymes and metabolic pathways. Perspect Sci. 2016;9:24–32.
21. Kandlinger F, Plach MG, Merkl R. AGeNNT: annotation of enzyme families by means of refined neighborhood networks. BMC Bioinformatics. 2017;18:274.
22. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta. 1975;405:442–51.

23. Kilmartin JR, Maher MJ, Krusong K, Noble CJ, Hanson GR, Bernhardt PV, Riley MJ, Kappler U. Insights into structure and function of the active site of SoxAX cytochromes. J Biol Chem. 2011;286:24872–81.
24. Bagchi A, Ghosh TC. A structural study towards the understanding of the interactions of SoxY, SoxZ, and SoxB, leading to the oxidation of sulfur anions via the novel global sulfur oxidizing (*sox*) operon. Biochem Biophys Res Commun. 2005;335:609–15.
25. Lawrence J. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev. 1999;9:642–8.
26. Merkl R, Wiezer A. GO4genome: a prokaryotic phylogeny based on genome organization. J Mol Evol. 2009;68:550–62.
27. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct. 2000;29:105–53.
28. Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. Chem Rev. 2016;116:4884–909.
29. Chojnacki S, Cowley A, Lee J, Foix A, Lopez R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic Acids Res. 2017;45:W550–W3.
30. Python Software Foundation. Python Language Reference, version 2.7 [http://www.python.org]. Accessed 5 Dec 2019.
31. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci U S A. 2000;97:6652–7.
32. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27:431–2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.