Mobile Genetic Elements

ISSN: (Print) 2159-256X (Online) Journal homepage: http://www.tandfonline.com/loi/kmge20

Ordered orthology as a tool in prokaryotic evolutionary inference

Sagi Snir

To cite this article: Sagi Snir (2016) Ordered orthology as a tool in prokaryotic evolutionary inference, Mobile Genetic Elements, 6:6, e1120576, DOI: 10.1080/2159256X.2015.1120576

To link to this article: http://dx.doi.org/10.1080/2159256X.2015.1120576

Accepted author version posted online: 30 Dec 2015.
Published online: 30 Dec 2015.

MOBILE GENETIC ELEMENTS
2016, VOL. 6, NO. 6, e1120576 (10 pages)
http://dx.doi.org/10.1080/2159256X.2015.1120576

Taylor & Francis
Taylor & Francis Group

REVIEW

Ordered orthology as a tool in prokaryotic evolutionary inference

Sagi Snir

Department of Evolutionary Biology, University of Haifa, Haifa, Israel

ABSTRACT
Molecular data is accumulated at exponentially increasing pace. This deluge of information should have brought us closer to resolving one of the most fundamental issues in biology - deciphering the history of life on Earth. So far, however, this abundance of data only seems to blur our understanding of the problem. This is largely due to horizontal gene transfer (HGT), the transfer of genetic material between evolutionarily unrelated organisms that transforms the prokaryotic tree into a network of relationships. Recently, we developed a method to infer evolutionary relationships among closely related species where the conventional evolutionary markers do not provide a strong enough signal. The method relies on the loss of synteny, gene order conservation among species that provides a stronger signal, sufficient to classify even strains of a given species. Here we elaborate on this method and suggest further uses of it in the context of detecting HGT events and genome architecture.

ARTICLE HISTORY
Received 10 March 2015
Revised 27 October 2015
Accepted 10 November 2015

KEYWORDS
genome architecture; horizontal gene transfer; phylogenetics; synteny; tree of life

Introduction

It is generally believed that the exponentially accumulating genetic molecular data will bring us closer to resolving one of the most fundamental issues in biology - understanding the history of life on Earth. The advent of High-throughput sequencing brought us even closer to achieving this goal by allowing the integration of genome analysis and systematic studies, an area called Phylogenomics.[1-3] So far, however, this abundance of data seems only to make reaching that goal harder than initially thought. This is largely due to horizontal gene transfer (HGT), the passage of genetic material between organisms not through lineal descent,[4,5] which, to a large extent, is mediated by viruses (bacteriophages), plasmids, transposons and other mobile elements. Collectively, these elements constitute the mobilome, that is, the enormous aggregate of genetic entities for which horizontal transfer is the dominant mode of dissemination.[6,7] Evolution in light of HGT tangles the traditional universal Tree of Life, turning it into a network of relationships.[8-11] Estimates of the fraction of genes that have undergone HGT vary widely with some as high as 99%. See e.g.[12,13] and references therein.

Despite the above, the belief in a single, underlying species tree still attracts efforts to overcome these confounding histories.[14-16] In particular there is ample evidence that a strong tree-like signal can be extracted, even in the presence of extensive HGT.[17-19] As a result, the bacterial phylogeny is usually inferred from genes that are thought to be immune to HGT, typically rRNA genes. Notwithstanding, even such genes are subjected to HGT, obfuscating the central trend of evolutionary relationships.[20-22] Moreover, as these genes are highly conserved, the amount of evolutionary signal they provide falls short for reliable classification within a genus or even a family. Alternatively, it was shown that gene order conservation among related genomes, denoted as synteny,[23,24] or rather the loss thereof is more informative and is very strongly correlated with amino acid distance.[25]

The study of HGT, or in general of prokaryotic evolution of which HGT is undoubtedly among the major factors, is of prime importance from several aspects. First, from medical perspective, HGT plays a major role in the emergence of new human diseases, as well as promoting the spread of antibiotic resistance in bacteria species.[26,27] From the fundamental, evolutionary standpoint, HGT links distant branches in the tree

CONTACT Sagi Snir sagi.snir@gmail.com Department of Evolutionary and Environmental Biology, University of Haifa, Mount Carmel, Haifa 31905, Israel.

of life, introducing immense variability to the gene repertoire of organisms. Genetically, the mobilome is an important, if not the primary, source of new genes that are acquired by bacteria and archaea and often result in adaptations to new environments and conditions.[28] Recent advances of comparative genomics and especially metagenomics indicate that the genetic complexity of the mobilome is vast and exceeds by several orders of magnitude the complexity of the set of conserved genes that are mostly vertically inherited.[29]

In this mini survey, we describe our 2 recent works aimed at harnessing loss of synteny for the goal of constructing the species tree depicting the central trend of microbial evolution.[30] We also expand and discuss other applications of this evolutionary footprints such as identification of HGT between closely related organisms,[31] and tracing exceptional genome architecture. We note that each of the above applications has been pursued to some degree in the past. Phylogenetic reconstruction based on gene order was proposed more than 2 decades back.[32,33] These works however deal with a different model of evolution that is assumed here, in which a genome undergoes types of operations denoted as genome rearrangement. They were motivated by early works pointing to the linkage between genome rearrangement events and evolutionary relatedness, starting with the classical work of Dobzhansky and Sturtevant on inversions in Drosophila chromosomes[34] and several others thereafter.[35] There has been a wealth of mathematical and biological extensions to the initial model, with more operations and finer algorithms and analysis (see Ref. [36-40] among many). To the best of our knowledge, practical works under this model were mostly confined to eukaryotes and HGT modeling was not included.

The task of detecting HGT has also been pursued extensively in the near past. This is commonly done by 2 orthogonal approaches. The phylogenetic approach[17,41] is accurate but relies on 2 very stringent requirements of accurate multiple sequence alignment and phylogenetic reconstruction. The sequence composition approach[42,43] relies on a strong enough signal in the DNA/protein sequence and is mainly confined to recent events due to amelioration. Finally, the wealth of the prokaryotic world never stops to surprise us with novel genome architectures,[25,44] in particular with new strains that are increasingly and constantly being sequenced. However, unifying these 3 tasks under a single umbrella based on a well-defined evolutionary process is novel and appealing.

SI based phylogenies

We now portray our synteny based phylogenetic method from.[30] We start with minimal necessary formal definitions. A genome is a sequence of genes and each gene is a sequence of DNA letters. That is, our view of a genome is at a resolution of genes, and of a gene at a resolution of nucleotides (Fig. 1A).

The *k-neighborhood* of a gene in a genome is the set of genes at distance at most k from it along the genome (i.e. at most k genes upstream or downstream). The conservation of gene order between 2 genomes is called synteny. Consider a gene common to 2 genomes. Then the k *synteny* index (k-SI) of that gene, or just SI when it is clear from the context, is the number of common genes in the 2 k neighborhoods of this gene in both genomes. We remark that in cases of circular genomes, a genome is broken arbitrarily at some location and the k neighborhood should be taken accordingly (i.e., circularly).

For the sake of completeness, if a gene is present only in one genome, we count its SI as zero. See Figure 1B for illustration.

A genome undergoes events of gene gain and loss in which genes are added or removed respectively. These events produce variation over the gene repertoire of the various genomes. A horizontal gene transfer (HGT) is the event in which a gene of a donor genome, is inserted at some position in another, recipient, genome. Given sufficient evolutionary time, the original ancestral homolog of the acquired gene, if such existed, will have either been replaced by the newly arrived gene, or will diversify, and so the 2 will not be identical. Because a genome may be viewed as a sequence of genes (see Fig. 2), the new gene will nearly always be integrated between 2 genes, unless integrated at the edge of a linear chromosome.

In the event of HGT, the chance that the gene maintains in the recipient genome its old k-neighborhood (that is, the k- neighborhood from the donor), or even part of it, is very small (assuming k is significantly smaller than the genome size). This means that the gene must be inserted at the same location it has in the donor genome. The above can be extended to the case of HGT of operons or gene clusters, in which a sequence of neighboring genes with a similar or

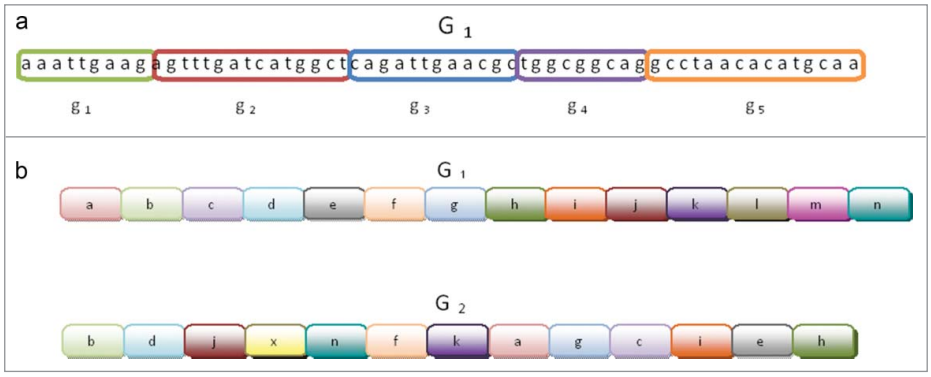

Figure 1. (A) A genome is viewed as a sequence of genes while a gene is a sequence of nucleotides. (B) Comparing G1 with G2 for k = 3, we get that the SI of gene g is 3, for x - 0, and for ` - 0. Figure was taken from.[30]

related function, located next to each other in the genome, are being copied. In such a case, only the genes at the 2 ends of the operons will have low SI while genes at the center will maintain their same neighborhood (of course this depends on the size of the operon and the neighborhood). Therefore, low SI of a specific gene, alludes to the possibility that the gene has undergone HGT. On a whole genome level, the more HGT a genome undergoes, the lesser its similarity, in terms of gene order, to related genomes. Hence, we expect that genomes which diverged long ago will exhibit small synteny to each other. In this case, we cannot use low SI of a certain gene as an indication for HGT of that gene, since most gene will exhibit quite low SI. However, we can use the SI to measure distances between the genomes exposed to high

HGT activity. Note that SI is defined only to genes that coexist in the 2 genomes. As genomes are more divergent from each other, such common genes are rarer. Hence for a gene found in only a single genome, we define its SI to be zero. We seek a measure that will consider the SI of all genes in the genome. We therefore take the average k-SI between the 2 genomes, denoted as SI. For two identical genomes, we have SI = 1 and for 2 genomes with disjoint sets of genes SI = 0. The SI therefore gives us a measure of similarity between pairs of species which we can use to construct evolutionary trees over the whole set of species. This property is attractive in particular for the organisms we investigate as they are subjected to heavy HGT activity resulting in different histories for different sets of genes. Therefore a method considering the aggregate set of genes is required. It is important to note here that the SI for a gene is not binary, i.e. either 1 or 0, as a result of being transferred or not. Genes can have SI values that range anywhere between 1 and 0 as a result of either being transferred with part of their original neighborhood, or of having kept their original neighborhood, but that neighborhood itself has been affected by additional HGT events.

When applying SI between all species, we obtain a similarity measure between the set of species. We can use this measure for phylogenetic reconstruction if we convert it to distances between the species. Hence we define a distances matrix [D] = 1 SI as our distance metric, where the subtraction operation is done for every 2 genomes in our taxa set. Note that every entry in D is between zero and one. Once we have the distance matrix D, we can then use it to reconstruct a phylogeny.

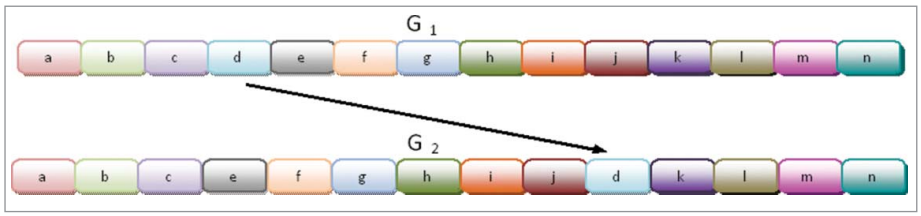

Figure 2. Gene d was transferred from Donor species G1 to recipient species G2. Figure was taken from.[30]

Distance based phylogenetic reconstruction methods, receive as input a symmetric dissimilarity matrix, representing dissimilarities between the taxa set under study, and return a tree over the taxa set. This tree should resemble optimally the input distance matrix. In our application we used the neighbor-joining (NJ)[45] algorithm implemented in the Phylip package.[46] We denoted the procedure illustrated above as Phylo SI. An immediate question that arises is how to set an efficient value for k (i.e., the neighborhood size) that obtains the maximum separability between genomes of different evolutionary distances. That is we seek such a k that will maximize the variance between the entries in the distance matrix.

Next, we turned to gauge the suitability of the SI method to phylogenetics reconstruction. We elaborate on this. The underlying assumption of our HGT model is that genes are acquired independently and integrated randomly in the genomes. The goal is to build a distance matrix from the average SI between all pairs of genomes, that reconstructs the tree edge lengths. We simulated genome evolution along a tree according to the same principles above (i.e. edge length signifies evolutionary distance). The resulted genomes at the leaves were given as input to the reconstruction methods. We compared the SI method to other genome wide reconstruction methods. We start with the traditional well established approach of genome rearrangement (GR),[32,37,38,47] that defines the distance between genomes as the number of operations required to transform one genome to the other. The original GR model requires genomes over the same gene set, so we restricted the simulation only to HGT events between the genomes simulated. The GR implementation software chosen was GRAPPA.[47] As algorithms for GR are very heavy, the study was restricted to unrealistic tree size of 10 species and tiny genomes of up to 80 genes. We used the Robinson-Foulds (RF) Symmetric Difference[48] tree metric to measure distance of the reconstructed tree to the originating model tree. Accuracy (RF distance) and running times were measured as a function of genome size. We found that the SI method is at least as good (accurate) as GRAPPA, regardless of the rate of HGT, and the advantage grows with the size of the genome. Moreover, running times of GRAPPA became prohibitive for genome sizes larger than 80 genes. Also, GRAPPA's running times grow exponentially with both the size of the genome and the rate of HGT. This limitation makes GRAPPA impractical for large scale analyses of hundreds of species.

Two other whole genome based approaches that were suggested in the past are Gene content (GC)[49] and Directed gene pairs (DP).[50,51] Under GC, the relative number of shared genes is counted. DP counts the relative number of ordered, uninterrupted gene pairs that are present in both genomes. This measure reects the degree of gene order similarity between the genomes. We found that the DP approach is significantly inferior to SI and GC. The GC approach may appear to be superior to the SI however this occurs only at unusual cases of very high rate of gene loss events compare to gene gain. Their consequence is a sharp reduction in the genome size up to the degree of genome vanishing. Needless to say that such cases are not found in nature as our real data and other studies suggest.[7] In order to apply Phylo SI to real microbial genomic data, we chose to use RefSeq annotation[52] as it offers an easy implementation of the algorithm. Being aware of RefSeq limitations such as very partial genome coverage (very few genes being identified) and hence also partial orthology mapping, we expected some inaccuracies in the output.

We applied our method to a set of 89 arbitrarily chosen complete bacterial genomes. We set $k = 10$ and used the neighbor joining algorithm for tree construction. The resulting tree is shown in Figure 3 with major phyla color-coded.

Although the tree contains several branches that correspond to known phyla [e.g., Firmiicutes (orange), Actinobacteria (pink), and Cyanobacteria (green)], and correctly places many sister taxa, it nevertheless has several inconsistencies with known taxonomic relations of bacteria. One large clade (marked in blue), divided into 2 sub-clades contains only gammaproteobacteria, but some gammaproteobacteria branch deeper with Bordetella which is a β-proteobacterium. Failure to separate these closely related classes is common in both genome-based analyses[53] and rRNA-based analysis.[54] However there are also more extreme cases such as members of the Aqui_cae and Thermotogae phyla that fall within the betaproteobacteria branch (cyan). As can be also seen from the figure, there is an uncolored clade (on the top) containing taxa from unrelated groups. A closer examination of the genomes in that clade, reveals that most of them contain only a handful of annotated genes under RefSeq, with 14 genes for the Alicyclobacillus

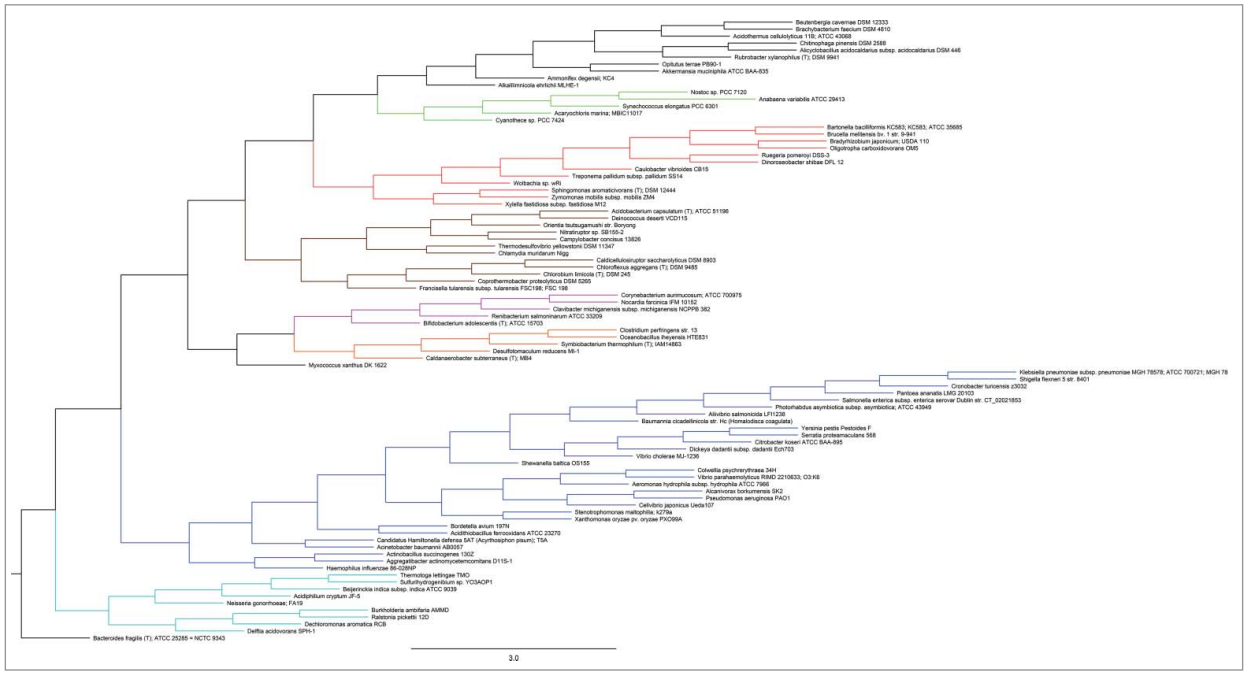

Figure 3. The SI tree on 89 microbial organisms. The tree was constructed by Neighbor Joining from pairwise distances [D] = 1 SI with k = 10. The tree is, by construction, fully resolved (86 internal branches). Figure is taken from.[30]

acidocaldarius as an extreme case. Needless to say, any reliable inference is impossible in such cases. Despite the above drawbacks, we compared this tree to 3 other trees (over the same taxa set) obtained by well accepted methods: the tree of life (TOL),[55] a 16S-rRNA based tree constructed using MaximumLikelihood from aligned sequences extracted from the RDP database,[56] and finally, the tree constructed with AMPHORA suit.[57] Additionally to the RF tree distance that is very strict, up to being uninformative, we added the quartet _t (Q_t) similarity measure[58] that counts how many quartets (4-taxa sets) have the same topology under both trees. The similarity of the SI tree to all these trees is very significant (14% and 66% for RF and Q_t respectively) where most differences are due to the poorly annotated genomes.

To alleviate the effect of partial coverage, we applied some filtering procedure, taking only genomes with at least 500 annotated genes to the species set. The similarity (tree is not shown) to the other 3 trees, TOL, 16S, and AMPHORA, is substantially higher. Also the biological soundness of the tree is much higher with the phyla Actinobacteria, Cyanobacteria both forming monophyletic clades, and the classes Gammaproteobacteria and Alphaproteobacteria being almost monophyletic, with the exception of Francisella tularensis and Orientia tsutsugamushi respectively. These two exceptional taxa, have genomes spread with repetitive elements: O. tsutsugamushi has nearly 4200 identical repeats of more than 200 bp in size, which account for over 37% of its genome,[59] whereas F. tularensis has 50 copies of the transposon ISFtu1 and a duplicated region of 33.9 kb.[60] The presence of a large number of repetitive elements can generate homologous recombination events and randomize gene order within the genome, especially in intracellular pathogens, such as F. tularensis and O. tsutsugamushi that tend to have very relaxed selection pressures on genomic changes and mutations alike. This randomization of gene order probably obscures the synteny-based phylogenetic signal in these species. On the one hand, this is a limitation of the Phylo SI method. On the other hand this makes the method particularly useful for identifying such unusual genomes, possibly alluding to outstanding evolutionary events in the history of the organism. We return to this later in the article.

Our second benchmark for the SI based method was to test it on a less diverged bacterial class that have been well studied before. This analysis demonstrates the advantage of Phylo SI over the conventional methods analyzing ubiquitous, highly conserved genes such as the best known phylogenetic marker, the 16S-rRNA, that does not provide sufficient taxonomic

resolution. We chose a set of 45 Alphaproteobacterial species. As before, we also took the TOL55 and the 16S-rRNA using RDP. Both the 16S-rRNA based and the TOL trees contain unresolved branches, especially more recently diverged clades within the trees (e.g. the genus Brucella). Also, in terms of branch support, the SI tree was superior to the 16S-rRNA tree by preserving 32 versus 26 branches (20% more). These may be the result of confounding evolutionary histories for different genes (TOL) due to horizontal gene transfer or to the lack of evolutionary signal due to insufficient number of substitutions (16S-rRNA and TOL). In regard to the biological context, all trees recovered well the deep relationships, such as those between families and orders (e.g., Rickettsiales, Rhizobiaceae or Brucellaceae), and thus all methods were fairly accurate.

Nevertheless, some mismatches were found. In particular, the clade containing the genus Brucella is well-resolved by the SI tree but not in the 16S-rRNA based or TOL trees. This classification was compared to an established tree, based on sequences of multiple manually selected genes[61] and the 2 have the same topology for the *Brucella* species examined.

Further application of the synteny index approach

We now sketch 2 other, directly related, applications to the Phylo SI algorithm that was outlined above. Both rely on the synteny index as a marker for evolutionary closeness.

Detecting HGT between closely related species

In the previous section we described how we use the average SI as a dissimilarity measure between genomes to construct a tree over bacterial genomes. Since SI is defined for a single specific gene shared by 2 genomes, it can be exploited also to gene specific studies. Below we detail on a recent such study.[31] The above implies that between closely related species (and in particular strains of a species) where synteny in general is highly preserved, if a gene has exceptionally low SI, we might suspect it has undergone HGT. However, it can be noted that translocation or duplication events, where a gene changes its place or copies itself in a genome respectively, cause the same effect - the gene has replaced its neighborhood and hence exhibit low SI with respect to genomes where it is intact since speciation. While we can eliminate duplication events by simply discarding genes with few copies, we cannot, based on SI only, distinguish between HGT and translocation. Therefore, to distinguish a gene undergone HGT from translocations or duplications, we rely on the fact that a translocated (duplicated) gene has been in its hosting genome since its split from another genome, in contrast to a gene recently acquired through HGT. This implies that the translocated gene was subjected to nucleotides substitutions (point mutation) for the time period since its split from the other genome. Hence the induced distance between orthologous genes in 2 genomes, is proportional to the time since their divergence (split).

We now rely on a very basic evolutionary effect that was established by us, dubbed as Universal Pacemaker (UPM) of genome evolution.[62,63] The UPM principle states that along every lineage in the evolution of life, all genes change their mutation rate in unison, as if adhering to a universal (but lineage specific) pacemaker. The above implies the following fundamental property, denoted as constant relative mutability (CRM), which we exploit in this part and is a direct outcome of the UPM: For every 2 genes residing in a common genome, mutating at 2 (not necessarily constant) rates, the ratio between these 2 rates is (approximately) constant at all times.

The CRM property can be utilized to our task in the following way (a detailed, formal, description is found in[31]). If a gene has undergone a HGT between 2 species, then the evolutionary distance between these very species according to this gene (i.e., the number of mutations along that gene between these species) has shortened, proportionally to the time of the HGT. However, since the HGT is unknown, this short distance between the species cannot be associated with certainty to a HGT event. It can also be due to conserveness of that gene, or to the case that this gene has slowed its rate along these specific lineages (recall that the evolutionary tree is not known and in particular, the gene tree of the respected gene is substantially jumbled). Now the CRM property comes to play. It manifests that regardless of the characteristic rate of the HGT suspected gene, and even if it slowed down its rate, it maintains (approximately) the same ratio to all other gene rates along that lineage. Therefore, the following is done: An additional witness gene, and 2 additional reference organisms, are taken arbitrarily. Now, the rate ratio between the HGT suspected gene

and the witness gene is calculated at both organism pairs, the reference organisms, and original organisms, where the low SI was detected (it can be shown that this rate ratio is obtained by dividing the respected distances between the organism pair).

Now, we treat the rate ratio obtained at the reference organisms, as the expected ratio (or our null hypothesis). By the CRM hypothesis, the latter ratio between the 2 gene rates (that was measured across the distance between the reference organisms) is expected to prevail along all lineages and between any 2 organisms and, in particular, between the original organisms. Hence we can use any hypotheses testing tool (e.g. _2) to test if the deviation in the rate ratio is below some threshold p-value and report the suspected gene as a putative HGT in the genome it had both very low SI and exceptional short distance simultaneously.

The above direction is liable to several pitfalls. Primarily, CRM is very loose as the UPM is substantially over-dispersed.[64] Also, the witness gene may have undergone HGT as well, impairing our ability to infer HGT with confidence. A possible remedy for this is to use a multitude of reference organisms and witness genes and apply another statistical or manual test over the aggregate results. Our first work in this direction31 of using 2 separate evolutionary footprints to detect HGT was rather a proof of concept in which several genes were detected and verified. Several follow-up works, with more stringent statistics and more comprehensive scrutiny of a wider gene and genome sets, are underway.

Identifying exceptional genome architectures

Genome architecture is defined as the totality of non-random arrangements of functional elements in the genome."[65] One obvious example of this non-randomness is gene synteny. Genome similarity between related species is a strong evidence that loss of synteny is tightly correlated with amino acid distance,[25] supporting our phylogenetic approach. Our results stand in agreement with these findings, notwithstanding we can exploit this fact to hunt exceptional genome architecture. Species or clades that are "dislocated" in the SI tree, that is, stand in substantial disagreement to their phylogenetic placement according to other markers, may hint to exceptional genome architecture with respect to their close relatives, as we already show above for the case of Francisella tularensis and Orientia tsutsugamushi. SI trees can be built over various sets of organisms and be contrasted with accepted trees. Misplaced" species (denoted as rogue taxa[66]) in the SI tree will be identified. Such inspection can be done manually (as we have done for the 45 alphaproteobacteria) or automatically when the trees are too big. The approach to single out such rogue taxa is done using the maximum agreement subtree (MAST) approach (some care must be taken as these subsets are not unique as we evidenced in another context[62]). A synteny profile can be constructed that will represent a group of related organisms to which rogue taxa can be compared. Specific genes or genomic regions in these rogue taxa are then identified and should point at a general pattern across several such organisms. These observations can lead to investigate the respective genomes in detail and, like any unusual phenomena, have the potential to elucidate general principles of the evolution of genome architecture.

Concluding remarks

In this review we elaborated on a new direction in the study of microbial evolution that is based on genome synteny - gene order conservation, as a tool to track evolving genomes. We described 3 possible applications based on recent works of ourselves, where the first deals with microbial phylogenetics,[30] and the second with detecting horizontal gene transfer (HGT).31 We believe this direction is promising as it augments a new source of information to the systematist toolbox that has not been fully exploited. Combining it with other sources of information e.g., single nucleotide mutations, as we describe in our HGT detection procedure, can yield further interesting applications.

Both the phylogenetic and the HGT detection applications, were mostly proof of concept and demonstrated the power of this direction, as well as its limitations. In the phylogenetic realm, our results on real data suggest that the power of our method is between closely related species, up to the level of genus. Above that level, our method reaches saturation where we arrive a state of saturation and we get $SI = 0$ between distant species. The latter calls for searching maximal subsets of taxa where no saturation is encountered. In the HGT detection realm, it would be beneficial to polish the CRM mechanism either by employing several witness genes, reference organisms, and an

insightful correction for multiple hypothesis testing, to account for the multilevel tests performed. As this direction relies heavily on accurate orthologous gene identification, we predict it will be involved in activity to this effect as well. A dedicated orthology data base such as COGs[67] or EggNOG,[68] however with gene order incorporated is a natural byproduct.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This research was partially supported by the Israeli Science Foundation (grant ISF 1852/14).

References

[1] Eisen JA, Fraser CM. Phylogenomics: Intersection of evolution and genomics. Science 2003; 300(5626):1706-7; PMID: 12805538; http://dx.doi.org/10.1126/science.1086292

[2] Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 2005; 6(5):361-75; PMID:15861208; http://dx.doi.org/10.1038/nrg1603

[3] Zhulin IB. It is computation time for bacteriology! J Bacteriol 2009; 191(1):20-22; PMID:18978045; http://dx.doi.org/10.1128/JB.01491-08

[4] Doolittle WF. Phylogenetic classi_cation and the universal tree. Science 1999; 284(5423):2124-9; PMID:10381871; http://dx.doi.org/10.1126/science.284.5423.2124

[5] Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405(6784):299-304; PMID:10830951; http://dx.doi.org/10.1038/35012500

[6] Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Micro 2005; 3(9):722-32; http://dx.doi.org/10.1038/nrmicro1235

[7] Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucl Acids Res 2008; 36(21):6688-6719; PMID:18948295; http://dx.doi.org/10.1093/nar/gkn668

[8] Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 2002; 19 (12):2226-2238; PMID:12446813; http://dx.doi.org/10.1093/oxfordjournals.molbev.a004046

[9] Zhaxybayeva O, Lapierre P, Gogarten JP. Genome mosaicism and organismal lineages. Trends Genet 2004; 20:254-260; PMID:15109780; http://dx.doi.org/10.1016/j.tig.2004.03.009

[10] Peter Gogarten J, Townsend JRP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Micro 2005; 3(9):679-687; http://dx.doi.org/10.1038/nrmicro1204

[11] Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 2005; 5:33; PMID: 15913459; http://dx.doi.org/10.1186/1471-2148-5-33

[12] Dagan T, Martin W. The tree of one percent. Genome Biol 2006; 7(10):118; PMID:17081279; http://dx.doi.org/10.1186/gb-2006-7-10-118

[13] Galtier N, Daubin V. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci 2008; 363:4023-4029, ; PMID:18852109; http://dx.doi.org/10.1098/rstb.2008.0144

[14] Olga Zhaxybayeva JPG, Charlebois RL, Ford Doolittle W, Thane Papke R. Phylogenetic analyses of cyanobacterial genomes: Quanti_cation of horizontal gene transfer events. Genome Res 2006; 16(9):1099-1108; PMID: 16899658; http://dx.doi.org/10.1101/gr.5322306

[15] Abby SS, Tannier E, Gouy M, Daubin V. Lateral gene transfer as a support for the tree of life. Proc Natl Acad Sci 2012; 109(13):4962-4967, ; PMID:22416123; http://dx.doi.org/10.1073/pnas.1116871109

[16] Galtier N. A model of horizontal gene transfer and the bacterial phylogeny problem. Systematic Biol 2007; 56(4):633-642; http://dx.doi.org/10.1080/10635150701546231

[17] Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci USA 2005; 102:14332-14337; PMID:16176988; http://dx.doi.org/10.1073/pnas.0504068102

[18] Puigbo P, Wolf YI, Koonin EV. The tree and net components of prokaryote evolution. Genome Biol Evolution, 2010; 2:745-756; http://dx.doi.org/10.1093/gbe/evq062

[19] Koonin EV, Puigbo P, Wolf Y. Comparison of phylogenetic trees and search for a central trend in the forest of life. J Computational Biol 2011; 18(7):917-924; http://dx.doi.org/10.1089/cmb.2010.0185

[20] van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD. Discordant phylogenies within the rrn loci of rhizobia. J Bacteriol 2003; 185(10):2988-2998; PMID:12730157; http://dx.doi.org/10.1128/JB.185.10.2988-2998.2003

[21] Schouls LM, Schot CS, Jacobs JA. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 2003; 185(24):7241-7246; PMID:14645285; http://dx.doi.org/10.1128/JB.185.24.7241-7246.2003

[22] Dewhirst FE, Shen Z, Scimeca MS, Stokes LN, Boumenna T, Chen T, Paster BJ, Fox JG. Discordant 16S and 23S rRNA gene phylogenies for the Genus Helicobacter: Implications for phylogenetic inference and systematics. J. Bacteriol 2005; 187(17):6106-6118; PMID: 16109952; http://dx.doi.org/10.1128/JB.187.17.6106-6118.2005

[23] Engstrm PG, Ho Sui SJ, Drivenes Y, Becker TS, Lenhard B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 2007; 17 (12):1898-1908; PMID:17989259; http://dx.doi.org/10.1101/gr.6669607

[24] Sanko D, El-Mabrouk N. Genome rearrangement. In Jiang T, Xu Y, Zhang M, editors, Current topics in computational molecular biology. CRC Press 2002

[25] Novichkov PS, Wolf YI, Dubchak I, Koonin EV. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol 2009; 191(1):65-73; PMID:18978059; http://dx.doi.org/10.1128/JB.01237-08

[26] Pallen MJ, Wren BW. Bacterial pathogenomics. Nature 2007; 449(7164):835-842; PMID:17943120; http://dx.doi.org/10.1038/nature06248

[27] Donnenberg MS. Pathogenic strategies of enteric bacteria. Nature 2000; 406(6797):768-774; PMID:10963606; http://dx.doi.org/10.1038/35021212

[28] Daubin V, Ochman H. Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli. Genome Res 2004; 14(6):1036-1042; PMID:15173110; http://dx.doi.org/10.1101/gr.2231904

[29] Edwards RA, Rohwer F. Viral metagenomics. Nat. Rev. Microbiol 2005; 3:504510

[30] Shifman A, Ninyo N, Gophna U, Snir S. Phylo si: a new genome-wide approach for prokaryotic phylogeny. Nucleic Acids Res 2014; 42(4):2391-2404; PMID: 24243847; http://dx.doi.org/10.1093/nar/gkt1138

[31] Adato O, Ninyo N, Gophna U, Snir S. Detecting horizontal gene transfer between closely related taxa. Plos Comp Biol 2015; 11(10):e1004408; http://dx.doi.org/10.1371/journal.pcbi.1004408

[32] Sanko D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci 1992; 89(14):6575-6579; PMID:1631158; http://dx.doi.org/10.1073/pnas.89.14.6575

[33] Sanko D. Edit distance for genome comparison based on non-local operations. In Apostolico A, Crochemore M, Galil Z, Manber U, editors, Combinatorial Pattern Matching, volume 644 of Lecture Notes in Computer Science, pages 121-135. Springer Berlin Heidelberg 1992

[34] Dobzhansky Th, Sturtevant AH. Inversions in the chromosomes of drosophila pseudoobscura. Genetics 1938; 23(1):28-64; PMID:17246876

[35] Nadeau JH, Taylor BA. Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci 1984; 81(3):814-818; PMID:6583681; http://dx.doi.org/10.1073/pnas.81.3.814

[36] Piotr Berman and Hannenhalli, Sridhar. Fast sorting by reversal. In Hirschberg D, Myers G, editors, Combinatorial Pattern Matching, volume 1075 of Lecture Notes in Computer Science, pages 168-185. Springer Berlin Heidelberg, 1996

[37] Bafnabackgroundand V, Pevzner PA. Genome rearrangements and sorting by reversals. SIAM J. Comput 1996; 25 (2):272-289

[38] Hannenhalli S, Pevzner PA. Transforming men into mice (polynomial algorithm for genomic distance problem). In Proceedings of the 36th Annual Symposium on Foundations of Computer Science, FOCS '95, pages 1995:581-592

[39] Kececioglu JD, Sanko D. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. Algorithmica 1995; 13:180-210; PMID:NOT_FOUND; http://dx.doi.org/10.1007/BF01188586

[40] Moret BME, Wang LS, Warnow T, Wyman SK. New approaches for reconstructing phylogenies from gene order data. Bioinformatics 2001; 17(suppl 1):S165-S173; PMID:11473006; http://dx.doi.org/10.1093/bioinformatics/17.suppl_1.S165

[41] Jin G, Nakhleh L, Snir S, Tuller T. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol Biol Evol 2007; 24(1):324-37; PMID:17068107; http://dx.doi.org/10.1093/molbev/msl163

[42] Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 1997; 44(4):383-97; PMID:9089078; http://dx.doi.org/10.1007/PL00006158

[43] Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 2004; 36(7):760-6; PMID: 15208628; http://dx.doi.org/10.1038/ng1381

[44] Isambert H, Stein R. On the need for widespread horizontal gene transfers under genome size constraint. Biol Direct 2009; 4(1):28, ; PMID:19703318; http://dx.doi.org/10.1186/1745-6150-4-28

[45] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4(4):406-425; PMID:3447015

[46] Felsenstein J. PHYLIP - phylogenetic inference package, (version 3.2). Cladistics 1989; 5:164-166

[47] Moret BM, Wyman S, Bader DA, Warnow T, Yan M. A new implementation and detailed study of breakpoint analysis

[48] Robinson DR, Foulds LR. Comparison of phylogenetic trees. Mathematical Biosci 1981; 53:131-147; http://dx.doi.org/10.1016/0025-5564(81)90043-2

[49] Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet 1999; 21(1):108-110; PMID:9916801; http://dx.doi.org/10.1038/5052

[50] Huynen MA, Bork P. Measuring genome evolution. Proc Natl Acad Sci 1998; 95(11):5849-5856; PMID:9600883; http://dx.doi.org/10.1073/pnas.95.11.5849

[51] Korbel JO, Snel B, Huynen MA, Bork P. Shot: a web server for the construction of genome phylogenies. Trends Genetics 2002; 18(3):158-62; http://dx.doi.org/10.1016/S0168-9525(01)02597-5

[52] Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res 2005; 33(suppl):D501-504; PMID:15608248

[53] Uri Gophna W. Ford Doolittle, and Robert L. Charlebois. Weighted genome trees: Re_nements and applications. J Bacteriol 2005; 187(4):1305-16; PMID:15687194; http://dx.doi.org/10.1128/JB.187.4.1305-1316.2005

[54] Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW. Phylogeny of gammaproteobacteria. J Acteriol 2010; 192(9):2305-14; http://dx.doi.org/10.1128/JB.01480-09

[55] Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science 2006; 311(5765):1283-7; PMID: 16513982; http://dx.doi.org/10.1126/science.1123061

[56] Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, et al. The ribosomal database project: improved alignments and new tools for rrna analysis. Nucleic Acids Res 2009; 37(suppl 1):D141; PMID: 19004872

[57] Wu M, Alexandra J. Scott. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. Bioinformatics 2012; 28(7):1033-4; PMID:22332237; http://dx.doi.org/10.1093/bioinformatics/bts079

[58] Jorge FE. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. System Biol 1985; 34(2):193-200

[59] Larsson P, Oyston PC, Chain P, Chu MC, Duffield M, Fuxelius HH, Garcia E, Hälltorp G, Johansson D, Isherwood KE, et al. The complete genome sequence of francisella tularensis, the causative agent of tularemia. Nat Genet 2005; 37(2):153-9; PMID:15640799; http://dx.doi.org/10.1038/ng1499

[60] Cho NH, Kim HR, Lee JH, Kim SY, Kim J, Cha S, Kim SY, Darby AC, Fuxelius HH, Yin J, et al. The orientia tsutsugamushi genome reveals massive proliferation of conjugative type iv secretion system and hostcell interaction genes. Proc Natl Acad Sci 2007; 104 (19):7981-6; PMID: 17483455; http://dx.doi.org/10.1073/pnas.0611553-104

[61] Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PSG, Roberto FF, Hnath J, Brettin T, Keim P. Whole-genome-based phylogeny and divergence of the genus brucella. J Bacteriol 2009; 191(8):2864-2870; PMID:19201792; http://dx.doi.org/10.1128/JB.01581-08

[62] Snir S, Wolf YI, Koonin EV. Universal pacemaker of genome evolution. PLoS Comput Biol 2012; 8(11): e1002785; PMID:23209393; http://dx.doi.org/10.1371/journal.pcbi.1002785

[63] Muers M. Evolution: Genomic pacemakers or ticking clocks?. Nat Rev Genet 2013; 14:81; PMID:23247404; http://dx.doi.org/10.1038/nrg3410

[64] Wolf YI, Snir S, Koonin EV. Stability along with extreme variability in core genome evolution. Genome Biol Evol 2013; 5(7):1393-402; PMID:23821522; http://dx.doi.org/10.1093/gbe/evt098

[65] Koonin EV. Evolution of genome architecture. Int J Biochem Cell Biol 2009; 41(2):298-306. Molecular and Cellular Evolution: A Celebration of the 200th Anniversary of the Birth of Charles Darwin

[66] Wilkinson M. Majority-rule reduced consensus trees and their use in bootstrapping. Mol Biol Evolution 1996; 13 (3):437-44; http://dx.doi.org/10.1093/oxfordjournals.molbev.a025604

[67] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classi_cation of proteins from complete genomes. Nucl Acids Res 2001; 29(1):22-8; PMID:11125040; http://dx.doi.org/10.1093/nar/29.1.22

[68] Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. Nucl Acids Res 2008; 36(suppl):D250-254