Research Article

Open Access CrossMark

Genomic taxonomy boost by lexical clustering

Abstract

In the post-genomic era, drawing inferences from multiple massive data sets is a ubiquitous challenge in the computational life sciences. Multiple sequence alignment has played a key role in genomics (and other "omics") as a means of summarizing and representing relationships between sequences. However, two problems with alignment-based strategies are apparent: the computational expense of constructing alignments and the sensitivity of subsequent analyses to alignment uncertainties. Here we present a novel alignment-free alternative. We use frequency profiles (or *n*-gram vectors) for sequence comparison, a method inspired by lexical statistics. Such profiles can be used to infer relationships between texts or between biological sequences, and we demonstrate that two statistical techniques–hierarchical clustering (HC) and no*n*-negative matrix factorization (NMF)–provide invaluable insights in both contexts. We present four case studies. First, we show that bigram frequency profiles can be used to reconstruct the ontology of 102,402 PubMed titles selected for their relevance to nine drugs and nine therapeutic proteins. Second, we apply the same methodology to classify 63 protein kinase coding DNA sequences into functional categories, based on trigram frequency profiles. The two major classes (*Tyr* vs *Ser/Thr*) are correctly identified. Third, and similarly, we show that *Alu* subfamilies can be identified in 58,122 *Alu* sequences, in perfect agreement with the accepted topology of the *Alu* phylogeny, again based only on trigram frequency profiles. Fourth, we clustered 8,885 human promoters using trigram frequency profiles for ab initio discovery of co-expression networks associated with disease. We demonstrate that "lexical" statistics offers a viable alignment-free approach to identifying and representing structural, functional and evolutionary relationships. We envision that our approach will be applicable to rapid and revealing comparison of whole individual genomes, and will be an important tool for analysis and correlation of "omics" data.

Volume 1 Issue 1 - 2014

Kosi Gramatikoff,[1] Sarah E Boyd,[2] Jeffrey W Smith,[1] Jonathan M Keith[3]

[1]Sanford-Burnham Medical Research Institute, Cancer Research Center, USA
[2]SBI-EMBL Australia, Monash University Clayton Campus, Australia
[3]School of Mathematical Sciences, Monash University, Australia

Correspondence: Kosi Gramatikoff, Sanford-Burnham Medical Research Institute, Cancer Research Center, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA,
Email kosi.gramatikoff@genlogica.com

Received: April 13, 2014 | **Published:** May 21, 2014

Summary

Multiple sequence alignment is a foundational technique in bioinformatics, and is often the first step in DNA and protein sequence analyses. However, it can be a slow step for genomic scale datasets, a problem that will only get worse as the sheer scale of biological sequence analyses continues to increase. Sequence alignment is also potentially inappropriate when there have been many small- and large-scale rearrangements among the sequences to be aligned, and subsequent analyses may be sensitive to uncertainties in the alignment. In this paper, we propose an alignment-free methodology for sequence comparison, based on *n*-gram frequency vectors, and demonstrate its ability to detect ontological relationships in biological literature and DNA sequence families (specifically kinases, *Alu* repeats and promoter sequences of co-expression networks). The methodology is versatile for clustering methods such as classical hierarchical clustering, as well as no*n*-negative matrix factorization. It is also highly efficient in terms of computational time and space requirements, and we foresee it becoming an indispensable tool in genomic sequence analysis.

Introduction

August Schleicher, Ernst Haeckel, and other 19th century linguists viewed language as a living system.[1] Darwin was influenced by this view and proposed that evolution of species and evolution of language are similar.[2] Several attempts have been made to introduce insights from linguistic theory into biology.[3–5] For example, Botstein and Cherry[6] proposed rules for general "molecular linguistics". Brendel et al.[7] used formal linguistic concepts to define a basic grammar for genes, based on the idea that mutating a piece of genetic information was similar to modifying words. However, the correspondence between biology and linguistics remains a matter of debate.[8,9]

Despite linguistic intuitions by Estoup-Zipf, Turing, Shannon and Gamow that now underpin information principles in biology,[10–12] the core sequence analysis methodologies of sequence alignment and phylogenetics are no*n*-linguistic.[13] However, a linguistic-like alternative – becoming known as *alignment-free* methodology -has been developing from attempts to overcome limitations of sequence-based alignment methods in phylogenomics.[14–16]

This methodology has several manifestations:

i. Graphical representations for visualization of DNA primary sequences based on visual maps of corresponding short DNA sequences.[17]

ii. Gene content and phylogenetic tree reconstruction.[18–20]

iii. Compression algorithms to measure relative information between sequences using Limpel-Ziv complexity.[21–23]

iv. Word-like or *n*-gram representation of DNA sequence and composition.[24–27]

In the 1940s Claude Shannon studied the information in natural language through the "*n*-gram" or "order" statistics. An *n*-gram is a set of *n* adjacent linguistic items, which may be letters or words. The *n*-gram statistics of language (or any particular corpus) are the frequencies with which each possible *n*-gram occurs. For random sequences, all *n*-grams are equally likely; to the extent that the language is structured and therefore predictable, the *n*-gram statistics will be highly no*n*-uniform. In some early works, the hope was expressed that sufficiently high order statistics would be able to fully capture the structure of language. However, Chomsky's[28] insistence that the structure of language is not a finite state dampened enthusiasm for the *n*-gram approach. It is believed that *n*-gram statistics pick up local rather than global (unbounded) dependences, and hence are unable to capture the structure of natural languages. Nonetheless, the use of *n*-gram-statistics is computationally less expensive, and therefore remains relevant to both linguistics and genomics. The application of *n*-gram statistics in genomics may be particularly useful for genomics, as the structure of the genome is possibly a finite state.

In the mid-1980s, *n*-grams were adapted for use in comparing gene sequences,[24] an approach typically called "alignment-free". In this approach, similarity among individual sequences is gauged by comparing the frequency of all *n*-grams.[14,29] These methods overcome a disadvantage of the classical Smith–Waterman alignment algorithm (used in BLAST and FASTA), which assumes conservation of contiguity between sequences. Consequently, Smith–Waterman alignment is confounded by sequence rearrangement, whereas alignment-free methods are "free" of this constraint. Alignment-free methods have been used to infer *hyponymic* (*kind_of*) or taxonomic relationships.[30] However, no application of alignment-free methods has been applied to meronymic (*part_of*) and cause-effect relationships. While taxonomic relations are a well-established area of research, investigations concerning meronymic relations of part-whole and causal relationships are relatively rare. This is surprising because this knowledge plays an important role in cognition, and cause-effect relationships affect all aspects of life. Here we show how a new alignment-free approach can be used to infer ontological relationships in language and DNA sequences. We gather large numbers of natural language (PubMed titles) or DNA strings into ontological groups based on prior annotations and sum their *n*-grams to build vectors that represent group-specific properties. The relationships between these vectors are then determined with hierarchical clustering (HC) and No*n*-negative Matrix Factorization (NMF).

The general utility of the strategy is illustrated with three examples:

i. Classification of biomedical literature.

ii. Classification of sequences of coding and no*n*-coding DNA.

iii. Elucidation of structure-function relationships among physiologically different co-expression networks. To the best of our knowledge, this is the first use of a DNA-"lexical" (*n*-gram) approach to compare gene co-regulation.

Results

The relationship mapping strategy involves three basic steps (Figure 1). First, sequences or segments of text are collected into groups, with each group defined by a common annotation. Second, the frequencies of *n*-grams derived from these sequences are tallied to produce a vector associated with each group. Third, the group vectors are clustered using hierarchical clustering (HC) or No*n*-negative Matrix Factorization (NMF). Thus, the approach reveals relationships among the original annotations through the exploration of *n*-grams representing the original text or DNA. While relationships among vectors can be discerned and displayed in many different ways, we chose to explore their relatedness with HC and NMF. HC is convenient and can be performed with widely available statistical packages. The statistical significance of individual clusters can also be estimated.[31,32] However, HC has the disadvantage that it imposes a stringent tree structure on the data. Moreover, the clustering produced by HC may vary depending on the distance metric used to assess similarity and on the "linkage" method (that is, on how distances between clusters are determined from the distances between their component vectors).

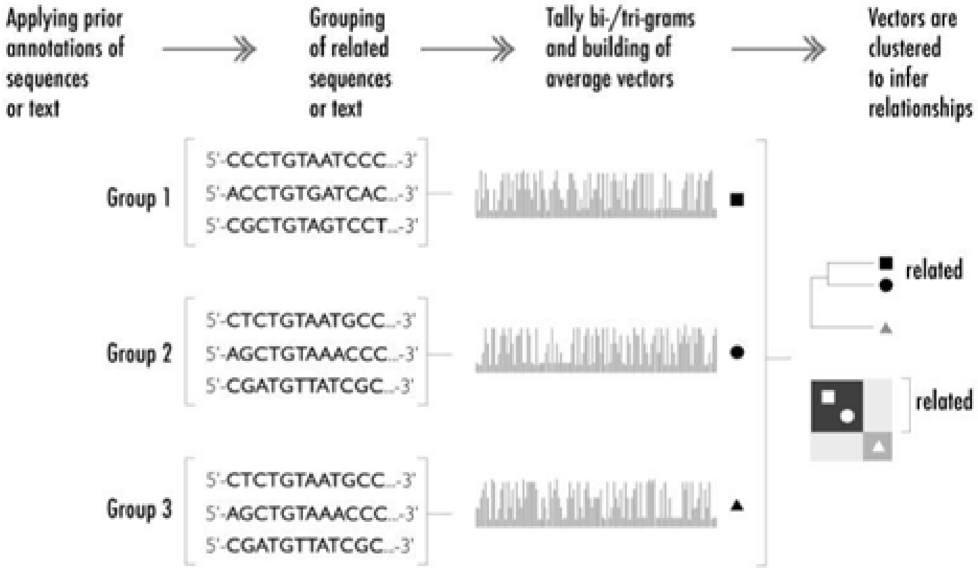

Figure 1 Method overview: schematic overview of the method with the three steps of data extraction, processing and interpretation. On the left, related groups of DNA sequences (or alternatively publication titles) are collated. The frequency of all possible trigrams (or bigrams for publication titles) is then tallied – labeled here with geometric shapes (square, circle, triangle). The relationships among the family representation vectors are mapped and compared through hierarchical clustering (HC) and No*n*-negative Matrix Factorization (NMF) (on the right).

In contrast, NMF produces a low dimensional approximation of a high-dimensional data matrix, in the form of no*n*-negative factors. The no*n*-negativity of these factors allows them to be interpreted as clustering of the data, where cluster assignments are not mutually exclusive. In this way, NMF can reveal hierarchical structure when it exists but does not force such structure on the data, and therefore has an advantage in exposing meaningful global hierarchy.[33] The robustness of NMF can be evaluated by a *cophenetic correlation coefficient* ρ.[34] NMF has been employed in the analysis of data where overlapping ontological structures may exist, such as in cancer class discovery and gene expression analysis[33,35] or in the biomedical literature.[36–38] However, when comparing the frequency vectors built through the *n*-gram decomposition of language or DNA, we would ideally like to combine both the ability of NMF to accurately and quantitatively identify overlapping structures, with the interpretability and visualization benefits of hierarchical techniques.

Inferring the ontology of publication titles using n-gram frequency vectors

As a first test of the ability of *n*-gram frequency vectors to indicate ontological structure, we attempted to recapitulate known ontological relationships in the biomedical literature. A corpus of biomedical publication titles with a known ontology was constructed as follows. Nine drugs were selected, comprising three functional groups of three drugs each (Figure 2). The functional groups contain drugs used to treat cardiovascular disease, neurodegenerative disease, and cancer. Within each of these three functional groups, three proteins were chosen based on their strong association with the respective diseases. The eighteen drug and protein names were then used as keywords to extract eighteen overlapping groups of publication titles from MEDLINE–a total of 102,482 unique titles. Note that there are no*n*-hierarchical ontological relationships in this data set, since the keywords can be clustered according to disease (cardio, neuro or cancer) or type (drug or protein) and neither classification is ontologically prior to the other.

The frequencies of all possible bigrams of English letters (26 x 26=676 bigrams) were evaluated for the titles corresponding to each of the eighteen keywords, with a total count of 8,934,899 bigrams. These bigram frequency vectors were then clustered using HC and NMF. Two distance metrics (Euclidean and correlation) and two linkage methods (average and complete linkage) were used for HC. The choice of distance metric had minimal impact on the resulting hierarchical clustering, whereas linkage method had a more significant effect (Supplementary Figure S1). All combinations of metric and linkage method were able to correctly identify three groups of proteins, and three groups of drugs. Moreover, the individual drugs and proteins were consistently and correctly allocated to these six groups, with the exception of one drug (bupivacaine) and one protein (DNMT). The drug bupivacaine was invariably clustered with the other neuro drugs, but only one combination (correlation metric, average linkage) found this membership to be significant at the 95% threshold. The protein DNMT was separated from the other cancer proteins by the cardio proteins in all but one combination (again, correlation metric with average linkage was the exception) and was not a member of a significant cluster for any combination. Interestingly, one possible explanation of DNMT not clustering with high significance, and appearing in both the cancer and cardio clusters comes from recent speculation that atherosclerosis is analogous to cancer, in that it involves similar phenotypic alterations and widespread hypomethylation in affected tissues.[39] Overall, the combination of correlation metric and average linkage method performed marginally better at detecting the six groups and these results are presented for comparison with NMF (Figure 2A).

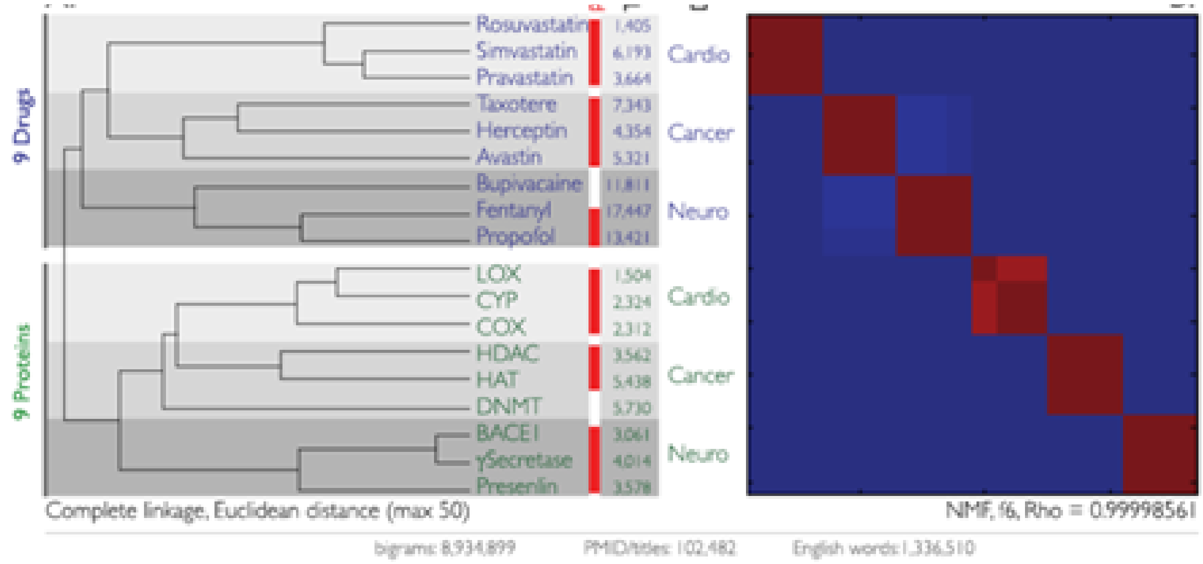

Figure 2 Comparison of language vector space via hierarchical clustering (HC) and No*n*-negative Matrix Factorization (NMF). A corpus of 102,482 publication titles was extracted from MEDLINE using nine keywords for drugs and nine for proteins. Frequency vectors have length 676 normalized counts corresponding to all English letter combinations (26 x 26). A total of 8,934,899 bigrams were clustered through HC (A, left) and NMF (B, right). The drug and protein vectors each self-organized into three isomorphic disease clusters (drugs in the top three clusters, purple labels; proteins in the bottom three clusters, red labels). These clusters are encoded Cardio, Cancer, Neuro for cardiovascular, cancer and neurological/CNS related diseases, respectively. A. Dendrogram of the language vectors based on Euclidean distance measure for the nine drugs and nine proteins. Three shades of gray illustrate the two topologically isomorphic clusters, which are subdivided in the three disease associations. Red bars indicate statistically significant clusters according to the approximately unbiased p-value

(alpha=0.95), as calculated by pvclust. B. NMF stability is shown as a heat-map after clustering of the same set of language-vectors. The NMF results in high cophenetic correlation coefficient rho (ρmax=0.999985561).

Citation: Gramatikoff K, Boyd SE, Smith JW, et al. Genomic taxonomy boost by lexical clustering. *J Investig Genomics.* 2014;1(1):13–25.
DOI: 10.15406/jig.2014.01.00004

In terms of detecting higher levels of ontology, the complete linkage method is superior in that it has correctly separated the drugs from the proteins (Figure 2A, left). The ability to identify this distinction based only on bigram frequencies is remarkable, and highlights the potential of the approach. The average linkage method, although less successful at detecting deeper levels of ontology, did cluster the cardio drugs with the cardio proteins, and the neuro drugs close to the neuro proteins (Supplementary Figure S1). This approach may therefore be influenced by the alternative high-level classification according to disease. We speculate that the average linkage method may be more successful at detecting low-level clusters, whereas complete linkage may be more appropriate for detecting deeper relationships.

A heat map for the NMF is shown in Figure 2B (right). NMF showed very good agreement with HC (correlation metric and average linkage method), identifying three groups of three drugs and three groups of three proteins, with no drug or protein misallocated, and NMF added no overlapping structures. Because of the almost perfect resolution among the eighteen vectors, NMF produced a cophenetic coefficient close to 1.0 (ρ_{max}). Note that NMF did not identify the deeper ontological distinction between drugs and proteins, or between diseases.

Inferring phylogeny of the human kinases genes with *n*-gram clustering

The *n*-gram clustering method was then tested for the ability to reconstruct the ontology of the human kinases, the complete catalogue of which is referred to as the *kinome*. Nine sets of paralogous genes were selected as representatives of the kinome (see Methods). These nine genes are of two types: receptor tyrosine kinases (RTKs) and serine/threonine kinases (STKs). The frequency count for all *n*-grams in each set of paralogs was transformed into an *n*-gram vector. For DNA we used trigrams instead of bigrams, so that each trigram vector has a length of 64 (4x4x4). In total, the nine kinase genes yielded 122,542 trigrams that were subsequently clustered with HC and NMF (Figure 3). For HC, two linkage methods (complete or average) and two distances (Euclidean or correlation) were applied in all combinations. Only the combination of complete linkage and Euclidean distance is shown in Figure 3; all four combinations are shown in Supplementary Figure S2.

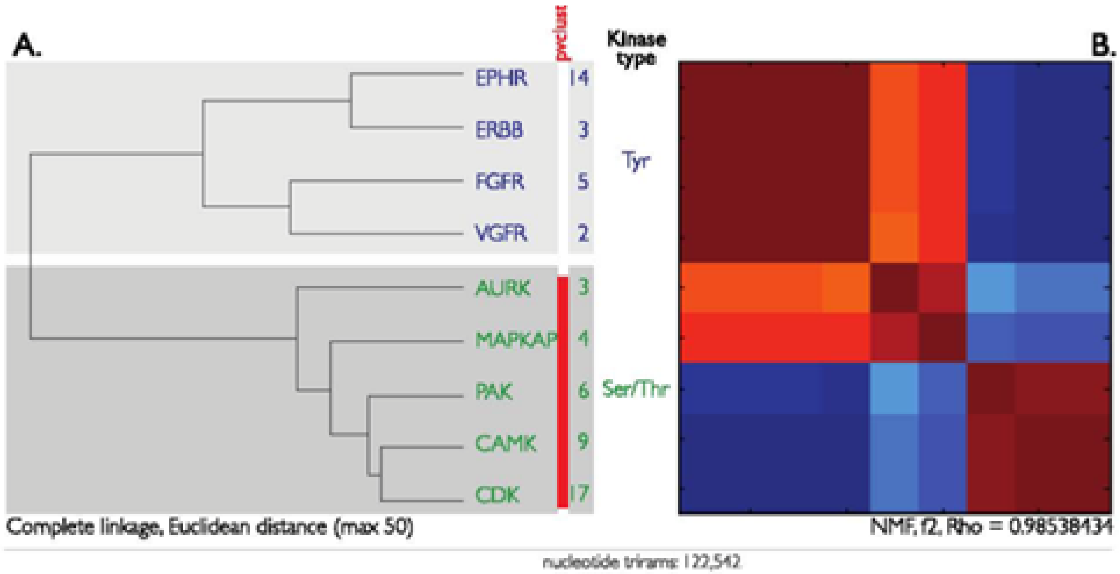

Figure 3 Kinase genes (coding sequence only) were extracted through BioMart service. Frequency vectors of length of 64 normalized counts for the four nucleotide-combinations (4x4x4), and a total of 122,542 trigrams were clustered through HC (A, left) and NMF (B, right). The two kinase types, receptor tyrosine kinase (RTKs) and serine/threonine (Ser/Thr) kinases, self-organized into separate clusters. The top cluster (purple labels) contains RTKs: 14xEPHRs (Eph receptor tyrosine kinase), 3xERBBs, v-erb-b2 erythroblastic leukemia viral oncogene homolog: 5xFGFRs, fibroblast growth factor receptor; 2xVGFRs, fms-related tyrosine kinase/vascular endothelial growth factor. The bottom cluster (red labels) contains Ser/Thr kinases: 3 x AURKs, Aurora family protein kinases; 4 x MAPKAPs, mitogen-activated protein kinase-activated protein kinases; 6 x PAKs, p21 protein (Cdc42/Rac)-activated kinases; 9 x CAMKs, calcium/calmodulin-dependent. Protein kinases; 17 x CDKs, Cyclin-dependent kinases. A. Dendrogram of DNA-vectors based on Euclidean distance measure for the 63 protein kinases. The red bar indicates the statistically significant cluster found with the approximately unbiased p-value (alpha=0.95), as calculated by pvclust. B. NMF is shown as a heat-map after clustering of the same set of DNA-vectors. The NMF results in high cophenetic correlation coefficient rho (ρmax = 0.98538434).

All four HC combinations distinguished RTKs from STKs, in agreement with classifications performed using sequence alignment. The HC method with Euclidean distance was also able to detect close evolutionary relationships between the pairs of RTKs, such as between the FGFR and VGFR families. These two RTK families originated early in metazoan evolution from a single gene that harbors a unique 7-intron code in the tyrosine kinase domain. Both FGFR and VGFR families of RTKs have characteristic arrays of immunogloblin (Ig) domains at the extracellular portion of the protein.[40,41]

NMF also distinguished RTKs from STKs with a high coefficient of correlation (ρ_{max}=0.98538434). However, it revealed an overlapping structure between the two types (Figure 3B). Previously, kinases were classified by comparing the DNA sequences of their catalytic domains[42] and by alignment-free methodology using amino acids.[43] Our *n*-gram approach allows an alignment-free analysis of the DNA sequences, and is not limited to the catalytic domains. Implicit consideration of multi-domain architectures is a valuable inclusion to complement other classification schemes because it draws attention to the non-hierarchical relationships among the kinases. Further detailed studies will characterize these relationships to determine if they correlate with particular structural domains or functional motifs.

Inferring phylogeny of Alu repeats with n-gram clustering

As a more challenging test, we applied *n*-gram clustering to the phylogeny of *Alu* repeats. *Alu* repeats are the most abundant mobile elements in the human genome and may contribute to phenotypic variation and disease.[44] This is a good illustration of the usefulness of the method because the vast number of *Alu* repeats precludes an alignment-based comparison across their full sequence. In fact, the current phylogenies are based on ~20 out of 130 diagnostic positions. We extracted 58,122 *Alu*s organized into nine distinct subfamilies. The selected *Alu*s included representatives of the major J, S, and Y classes.[45] The *n*-gram frequency counts for each *Alu* subfamily were calculated, transformed into *n*-gram vectors and then clustered by either HC or NMF. HC segregated the *n*-gram vectors into a phylogeny that almost perfectly matches the known *Alu* classification.[46] Moreover, the clustering was unaffected by the choice of linkage method or distance metric, although the significance of the clusters did vary (Supplementary Figure S3). Interestingly, NMF (coefficient of correlation ρ_{max}= 0.98557078) detects an overlapping in the *n*-gram structure of the older *Alu*J and *Alu*S classes, but not for the youngest *Alu*Yb and *Alu*Ya5 classes.

Previous age estimates of the *Alu* classes are based on sequence alignment using the Smith and Waterman algorithms (e.g. Allison-Yee and Needleman-Wunsch) combined with the Kimura's distance. These methods do not include insertions and deletions but take into consideration only the transitions and transversions.[46] It was determined that the average evolutionary distance and age of the *Alu*Jb and *Alu*Jo subfamilies are almost the same. In agreement with this sequence-based study, our HC clustering shows the lowest divergence between *Alu*Jb and *Alu*Jo (Figure 4A). When we compared our results with the standardized *Alu* nomenclature for the younger *Alu*Y subfamiles, the tri-gram sensitivity could be inferred. According to the *Alu* nomenclature, *Alu*Ya5 and *Alu*Yb89 differ by only three nucleotide changes.[45] Our results show that with the given set of 6,476 *Alu* repeats (*Alu*Ya5 and *Alu*Yb89), the tri-gram vectors for these two *Alu* lineages are indistinguishable by both HC and NMF. This implies that three diagnostic mutations may be insufficient to distinguish clusters when using tri-grams.

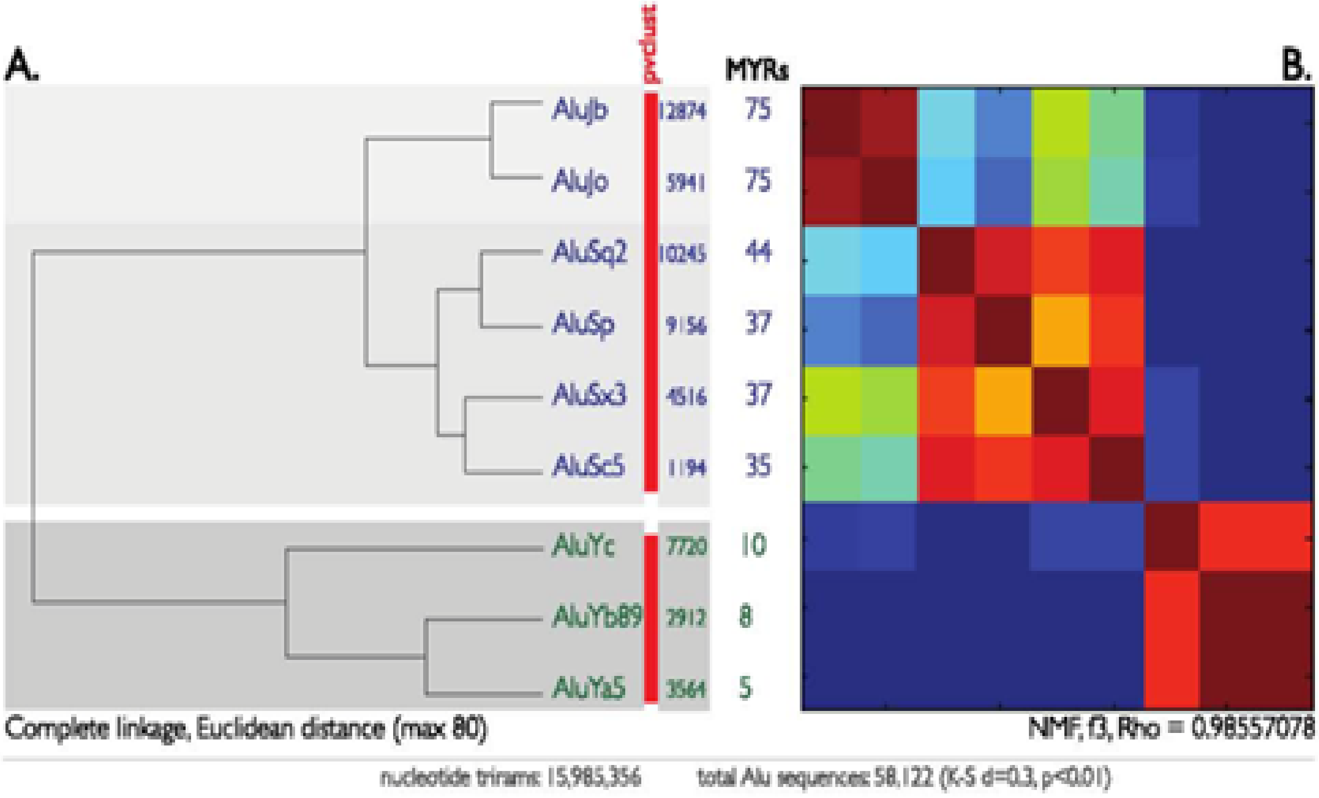

Figure 4 Repeats extracted using the UCSC genomic browser and BioMart service was analysed. Frequency vectors of length 64 normalized counts for the tri-gram nucleotide-combinations (4x4x4), with a total of 15,985,356 trigrams were clustered through HC (A, left) and NMF (B, right). Three evolutionary distant *Alu* subfamilies (*Alu*J, *Alu*S, and *Alu*Y) are self-organized into separate clusters. Evolutionary old (*Alu*Js) and middle-aged (*Alu*Ss) subfamilies are labeled in purple (A, top two clusters). Evolutionary young (*Alu*Ys) subfamilies are labeled with red labels (A, bottom cluster). The number of *Alu* sequences used for making the *n*-gram vectors is shown next to subclass name. The approximate evolutionary age of each *Alu* subfamily is shown in the middle (Million of Years, MYRs). A Dendrogram of DNA-vectors based on Euclidean distance measure for the *Alu*family. Red bars indicate statistically significant clusters according to the approximately unbiased p-value (alpha = 0.95), as calculated by pvclust. B. NMF is shown as a heat-map after clustering of the *Alu*-clustering of the *Alu*-vectors. The NMF results in high cophenetic correlation coefficient rho (ρmax = 0.98557078).

DNA-vectors based on Euclidean distance measure for the nine promoter networks involved in three diseases. Each network (n_) is labeled by a single gene name (n_Gene) to which co-expression of 300 non-redundant genes are correlated by COXPRESdb. The overlap of genes is shown in percentage on the dendrogram. The red bar indicates the statistically significant cluster according to the approximately unbiased p-value (alpha=0.95), as calculated by pvclust. B. NMF stability is shown as a heat-map after clustering of the DNA-vectors. NMF partitions the vectors comparable to the HC dendrogram with high cophenetic correlation coefficient rho (ρmax = 0.96569107). C. Example of gene overlap in two co-expression networks (n_MYL2 and n_MYH7) each centered at

Citation: Gramatikoff K, Boyd SE, Smith JW, et al. Genomic taxonomy boost by lexical clustering. *J Investig Genomics.* 2014;1(1):13–25.
DOI: 10.15406/jig.2014.01.00004

Inferring n-gram phylogeny of co-expression networks

We hypothesized that core RNA polymerase II promoters represent important ingredients of transcriptional co-regulation and therefore could be used for an *n*-gram unsupervised clustering of modeled co-expression networks. To test this hypothesis, we compared nine co-expression networks, each obtained by submitting nine gene names to COXPRESdb. The nine genes are the same as those used in the PubMed article title study above, and are thus grouped into three diseases, and annotated with "n_" preceeding the gene name to indicate that in this section, the focus is on networks (Figure 5). The *n*-gram frequency vectors for each network were then clustered using HC with average and complete linkage methods, and using Euclidean and correlation distance metrics. The choice of linkage method did not affect the clustering, although significances did vary (Supplementary Figure S4). Choice of distance metric was found to affect both clustering and significances. All combinations correctly identify the cardio co-expression networks n_COX2, n_MYL2, and n_MYH7 as a cluster. Using Euclidean distance, the cancer co-expression networks n_HDAC1, n_HAT1 and n_DNMT1 were clustered, as were the neuro networks n_PSEN2, n_APP and n_TUBB3. However, using correlation as a distance metric, n_HDAC1 and n_TUBB3 formed a significant cluster, and n_HAT1 and n_DNMT1 formed a non-significant cluster.

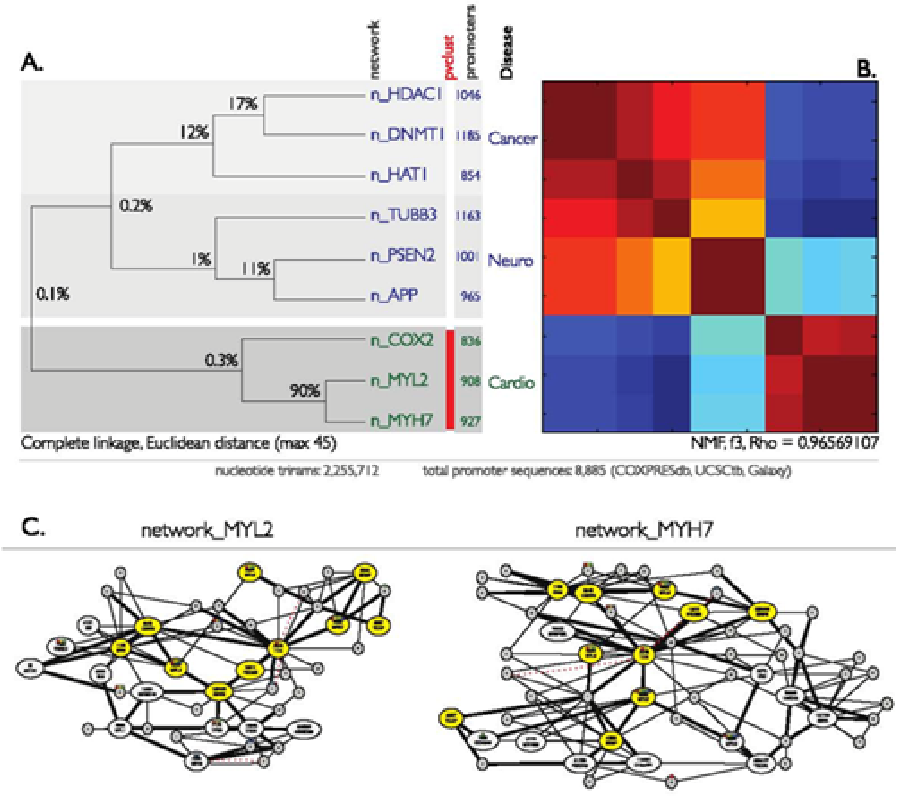

Figure 5 20 genes (text-labeled nodes) with an overlap of 50% (ten genes, yellow nodes). A corpus of 8,885 human promoters (utr) grouped into nine co-expressing networks was extracted through UCSC genomic browser. Frequency vectors with 64 normalized counts for the four nucleotide-combinations (4x4x4), or total 2,255,712 trigrams were clustered through HC (B, left) and NMF (C, right). The vectors for the nine promoter networks each self-organized into three disease clusters (Cancer, Neuro networks in the top six clusters, purple labels; Cardio networks in the bottom three clusters, red labels). These clusters are encoded Cardio, Cancer, Neuro for cardiovascular, cancer and neurological/CNS related diseases, respectively.

The clustering (with Euclidean metric) of the presenilin protein 2 (PSEN2) and amyloid precursor protein (APP) co-expression networks may be explained by their co-regulation in Alzheimer's disease.[47] Presenilin is postulated to regulate APP processing through its effects on gamma-secretase, an enzyme that cleaves APP. The co-clustering of the TUBB3 co-expression network, which encodes a class III member of the beta tubulin protein family, can be explained by its expression in neurons and involvement in neurogenesis, axon guidance and maintenance. The cardio networks n_COX2, n_MYL2, and n_MYH7 are also related, as follows. MYL2 gene encodes the regulatory light chain associated with cardiac myosin beta (or slow) heavy chain. MYH7 gene encodes the beta (or slow) heavy chain subunit of cardiac myosin. The clustering of these two genes can be explained by their co-regulation associated with cardiomyopathy.[48,49] In fact there is 90% overlap of the genes present in both the n_MYL2 and n_MYH7 co-expression networks, which explains their co-regulation and clustering. These results imply that the promoters for each disease category have similar *n*-gram vectors and perhaps similar structural motifs that determine their co-regulation.

To investigate the fact that some genes occur in more than one network (and therefore contribute identical *n*-gram signatures to each network in which they occur) we calculated the overlap of the clusters in percentage of genes (Figure 5A). This percentage correlates with the clusters, to the extent that percentage overlap might be sufficient to determine the relationships. For example, there is 90% overlap of the genes present in both the n_MYL2 and n_MYH7 co-expression networks, which explains the smallest distance in HC clustering and cluster indivisibility by NMF. In general, the clusters determined by HC are similar to the NMF solution ($\rho_{max} = 0.98095654$). NMF is able to detect no*n*-mutually exclusive clusters, and in particular strong overlap between the cancer and neuro structures exists. This no*n*-hierarchical structure was also apparent in the HC clustering with correlation metric, which identified clusters containing cancer and neuro genes, as noted above. NMF also reveals a weak overlap between the cardio and neuro structures.

Centering co-expression networks on a single gene may not be optimal for characterising the entire network, therefore we performed an ontology mapping of the nine co-expression networks via four separate ontologies (FunDO: Table 1; PANTHER, KEGG, TopoGSA, data not shown). The most consistent results were obtained via FunDO (Functional Disease Ontology) annotation.[50] FunDO takes a list of genes (Entrez gene IDs) and finds relevant diseases based on statistical analysis of the Disease Ontology annotation database. The result of the hypergeometric test used to assess the p-value of the enrichment, which is equivalent to Fisher's exact test (one-sided), is shown in Table 1. For each network we run two tests: with and without the top gene for which the network is named. The result shows that each of the nine co-expression networks is tightly associated with one of the three diseases (Cancer, Neuro [Alzheimer's disease], Cardio [Atherosclerosis and Myopathy]). The association to these diseases is statistically significant and there is only a slight drop in p-value when the major node is removed from the network. For example, the strongest association is of n_COX2 to Atherosclerosis (p-value=6.25E-28) followed by n_DNMT to Cancer (p-value=3.88E-26). Using the other three ontologies (PANTHER, KEGG, TopoGSA) also provided statistical evidence for the clustering of the network; however the terms in these ontologies were less focused on disease. For example, both PANTHER (Protein ANalysis THrough Evolutionary Relationships: www.pantherdb.org) and TopoGSA (Topology-based Gene Set Analysis) mapped the three networks n_HDAC1, n_DNMT1 and n_HAT1 to the same biological process: pyrimidine metabolism, nucleotide and nucleic acid metabolic process (p-values = 1.16E-16, 6.09E-25, 6.97E-13). The two networks n_MYL2 and n_MYH7 were also mapped to muscle development and cardiomyopathy (p-values=3.28E-35, 2.06E-37 respectively). Altogether, the ontological mapping confirmed that co-expression centered on a single gene can be used to characterize the disease relationship of correlated networks.

Table 1 FunDO statistics of the nine co-expression networks. Resulting p-values are given for testing both with the seed gene node included (+ top gene), and without the seed gene node (-top gene)

Network	FunDO term	p-value (+ top gene)	p-value(-top gene)
n_HDAC1	Cancer	7.59E-09	7.09E-09
n_DNMT1	Cancer	3.88E-26	3.35E-26
n_HAT1	Cancer	0.002796	0.002738
n_TUBB3	Alzheimer's disease	1.37E-08	1.31E-08
n_PSEN2	Alzheimer's disease	1.45E-03	1.43E-03
n_APP	Alzheimer's disease	1.31E-03	6.98E-03
n_COX2	Atherosclerosis	6.25E-28	1.42E-26
n_MYL2	Myopathy	4.24E-14	4.06E-14
n_MYH7	Myopathy	1.20E-12	3.36E-11

Discussion

We have shown how a new alignment-free methodology can be used to infer relationships in language and DNA sequences. For each ontological unit, an *n*-gram frequency vector was calculated and used to compare sequence pools gathered by a no*n*-sequence criterion. The originality of our approach is to infer relationships between sequences grouped by a functional criterion. Our study is structured in three conceptual parts. First we investigated how one could infer meronymic relationships in biomedical language. Second, we used the same approach on two pilot sets of well-characterized families of DNA sequences: kinase genes and *Alu* repeats. Third, we mapped relationships between co-regulated networks and their core promoters.

Linguistic-like features of DNA

It has been previously proposed[3,51,52] and commented on[8,53] that natural DNA sequences, and especially no*n*-coding DNAs, appear to have many statistical features in common with natural languages. In the original paper Mantegna et al.,[3] performed "linguistic tests" on DNA sequences, which are related to Zipf's distribution[10] and Shannon's information theory.[54] Their tests as well as our calculation (Supplementary Figure S5, Shannon's Capacity) seem to reveal significant differences between coding (CDS) and no*n*-coding regions (UTR) of natural DNA sequences. Statistical differences of coding and no*n*-coding DNA have been reported from as early as 1981, and are used even in routine methods for discrimination between the two,[55,56] in some cases forming the basis of these methods.[57] Using Shannon's Capacity for comparison of coding (CDS) and no*n*-coding (UTR) DNA shows little difference for short 'words' (2-5mers) and noticeable difference for longer 'words' (6-10mers) (Supplementary Figure S5). This may lead to the premature conclusion that Shannon's Capacity for short 'words' is indiscriminative. However, when short 'words' (3-mers, or codons) were used to compare 28 different species, remarkable evolutionary trend was observed, which could not simply be attributed to a global GC% change (Supplementary Figure S6 & S7). Using 3-mers (or codons), we also tested our alignment-free approach in a pa*n*-domain phylogeny for 44 species (grouped in five statistical brackets of codon counts distribution: www.kazusa.or.jp/codon) (Supplementary Figure S8 & S9). Our phylogeny relationships were essentially the same as the results by Ciccarelli et al.,[58] based on a conventional sequence alignment. These observations, however, have been strongly criticized by Konopka and Martindale.[59] Most of the observations made by Mantegna et al.[3] were explained by trivial consequences of uneven nucleotide frequencies. It was even concluded that Zipf-Shannon "linguistic" tests do not reveal any new biological information in either no*n*-coding or coding DNA.[60] Although these explanations remove any superficial evidence for this hypothesis, they do not rule out the existence of hidden "language" or lexical properties in DNA. Our results from using bigram vectors representing biological language ontologies and trigram vectors representing families of DNA sequences support the possibility of common principles of lexicostatistics between language and DNA. For example, the bigram vectors for cardiovascular drugs (statins) appear to distinguish both the chemical differences and/or timeline of development (Figure 2). This parallels to the idea of sequence phylogeny clearly shown in the case of the pa*n*-domain phylogeny (Supplementary Figure S8 &_S9) and the *Alu* repeats (Figure 4). Thus, if hidden DNA "language" exists, it may be discernable through multidimensional lexicostatistics and detectable through *n*-gram clustering.

Biomedical language as validation

Previously, the alignment-free similarity approach has been applied to infer relationships among English books of different genres, time period or authors.[26] It was shown that the approach clustered books that share the same author within the same genre. However, the nature of the relationship between books is not phylogenetic. Their semantic relationship is *hyponymic*, often referred as "is_a/type_of/*kind_of*" relationship. In their study, Sims et al.,[26] reasoned that the mapping of *hyponymic* relationships could provide an intuitive validation for their method in genomic comparisons. In contrast, we tested our alignment-free similarity approach for mapping simultaneously *hyponymic* and *meronymic-causal* relationships. The latter is often referred to as "overall/*part_of*" relationships. For example, Avastin, Herceptin, Fentanyl and Propofol are all *kind_of* drugs but the first two are *part_of* the cell division system and target cancer, while the second two are *part_of* the alkaloid receptor system and target neurons. NMF reveals both types of relationships simultaneously, because of its ability to represent no*n*-hierarchical ontologies.

What do n-gram frequency distances measure?

The *n*-gram relative frequency distance measure among language and genomic ontologies appears to detect meaningful relationships. It has been suggested that the tri-gram relative frequency values relate to DNA structures.[25,57] Several factors that influence DNA structures have been identified, e.g. dinucleotide stacking energies, curvature, superhelicity, methylation and other short oligonucleotide modifications, and DNA repair mechanisms.[61,62] For example, TpA is intrinsically less stable energetically than all other dinucleotides.[61] Flexibility of the TpA step is commonly associated with substantial DNA distortions. TpA models suggested that TpA sites could be important as nucleation sites for untwisting the DNA double helix. It appears that protei*n*-DNA complexes can exploit the reduced thermodynamic stability of the TpA base step. The TpA and ApT steps are conformationally incompatible causing a strain in the helix when juxtaposed, which can be relieved by unwinding the helix.[62]

DNA has at least two functions:

i. To provide special sequences for encoding gene products or for regulating transcription and

ii. To provide for genome replication and segregation. While the former requires some sequence specificity, the latter may be mostly DNA structure specific. In this vein, the relative frequency distances appear to assess and discriminate mostly local structure specificity.

Limitations and future applications

There are several limitations to our study. First, the tri-gram model may be too simplistic to capture domai*n*-specific relatedness. Thus, fixed and variable higher order Markov models may be needed. Second, our approach depends on the availability of accurate genomic annotation and known low-level ontological relationships among genes. Third, we explored promoters with the same length, which clearly is an over-simplification of gene regulation. Despite these limitations, our study provides evidence that our *n*-gram group statistics and clustering methods can potentially be used to study relationships between functionally distinct pools of genome sequences. As the *n*-gram distance can be calculated without sequence alignment and its computation for even a large number of sequences takes only seconds, *n*-gram distance can be extremely useful when we are faced with a large number of related sequences and may essentially be the only option in cases where there are sequences that are too diverged to be reasonably aligned.

Our approach is especially useful in two scenarios: where the sheer number of sequences to be compared renders alignment impractical or inefficient, and where the sequences to be compared are difficult or impossible to align.[63] With regard to the first scenario, recent massive delivery of genomic data is rapidly exceeding the capabilities of alignment methods and the problem will worsen within the next few decades, as massive re-sequencing of whole genomes and their active segments generates billions of closely related sequences.[64,65] The second scenario arises in sequences that are evolutionarily less constrained or where small-and large-scale sequence rearrangements have occurred. Wong and colleagues quantify the alignment uncertainty in genome-wide evolutionary analyses and reported that a staggering 46.2% of the 1,502 genes examined exhibit variation in the phylogeny produced that was dependent on the choice of alignment method.[15] But what about all the genes that are harder to align? One popular approach has been to exclude areas of uncertain alignment, and programs exist that do just that.[66] Filtering however, is unsuitable for studies where information from every site is potentially informative, or for studies of selection where rapidly evolving sites may be precisely those that are the most difficult to align.[63,64] The development of a new computational and statistical arsenal to account for the uncertainty stemming from sequence alignments is a much needed paradigm shift in the era of genome-scale analysis.[65]

We have demonstrated the main utility of the approach for analyzing co-regulatory networks. This analysis could be expanded to genome-wide classification of all possible co-regulatory networks that are associated with specific biological phenomena or processes. This could lead to building of large libraries of *n*-gram signatures that characterize networks of the cellular metabolism, cell-cycle progression, ageing and cell death. Combined with specific experimental platforms and assays the *n*-gram signatures could be validated and used for prediction at the organism level. We also limit the scope of this study to real-life examples that provide obvious and intuitive validation of the approach. A possible extension would be to evaluate the limit parameters and optimum parameter ranges for the methods. A previous study established limit parameters based on Shannon's entropy.[26] In Shannon's principles, however, the method of encoding *n*-gram vectors is based on the presupposition that the *n*-gram vectors are outcomes of a known random source -it is only the characteristics of that random source that determine the encoding, not the absolute characteristics of the *n*-gram vectors that are its outcomes, which is the more appropriate paradigm to consider.[67] However, a major difficulty with that approach is that the bounds of the Kolmogorov Complexity cannot be computed. Despite the problems with measurement, Kolmogorov Complexity and *n*-gram vector information are related in many ways. Cryptography, for example, attempts to take strings that have structure and make them appear random. The quality of a cryptographic system is related to the system's ability to raise the apparent complexity of the string, while keeping the actual complexity of the string relatively the same (within the bounds of the encryption algorithm). In other words, cryptography achieves its purpose by making a string appear to have a high Kolmogorov Complexity through the use of a difficult or impossible to guess algorithm or key. Thus, cryptoanalysis might be helpful to make better use of available information by constructing a composite model that incorporates *n*-gram frequency counts, pattern of preselected DNA words and or DNA dictionaries. Such a composite model might provide the statistical power of a high-order Markov model in a more versatile and effective fashion. Although much is known about statistical techniques for language recognition in theoretical Markov models, little is known about how well such models and techniques work for real-world examples of language and DNA.

A central aim of biology is to delineate the relationship between structure and function. Sequence complexity is inadequate to describe the phenomenon of function. Functions are like programming instructions that have been encrypted throughout evolution. A time has come to reconsider the dominant position of the sequence-based research strategy. We foresee the implementation of *n*-gram models over populations of functionally annotated sequences for mapping relationships as a future genomic strategy.

Methods

Biomedical corpus design and title extraction

We used finding-associated concepts with text analysis (FACTA), an online text search engine for MEDLINE abstracts that can quickly compute the association strengths between a query and different types of biomedical concepts based on their textual co-occurrence statistics.[68] While other similar systems exist, such as XplorMed,[69] MedlineR,[70] LitMiner[71] and Anni,[72] FACTA was chosen because of its ability to pre-index words and concepts, which results in fast, real-time responses. Two sets of biomedical concepts, namely "drug" and "protein" were examined and ranked through FACTA according to their frequencies of appearing in MEDLINE abstracts. As a result, the top nine "drugs" and "proteins" were grouped into three different disease categories: (i) cancer, (ii) cardiovascular and (iii) neurological. We used these nine drug names and nine protein names for text extraction of PubMed titles. Using simple keywords, we capitalize on the capacity of PubMed to automatically compare and map keywords from a user query to lists of pre-indexed terms (e.g. Medical Subject Headings, MeSH).[73,74] That is, if our query can be mapped to one or more MeSH concepts, PubMed will automatically add its MeSH term(s) to our original query. As a result, in addition to retrieving documents containing the query terms, PubMed also retrieves documents indexed with those MeSH terms. This technique boosts recall and is especially useful when the original keyword query has synonyms. In essence, the MeSH indexing, accomplished by trained indexers, offers a useful window into full text. We treat MeSH terms as high-quality hooks used to select important text segments in the full document. This gave us a general strategy for reducing biomedical documents to their core portions – the titles. After the submission of each drug and protein keyword query we took advantage of the "Send to function, which creates a file with CSV" format containing the titles associated with each individual keyword. The CSV files were imported to Microsoft Excel and titles were parsed in separate columns for further bigram calculation.

Extracting the corpus of kinase sequences

Sequence data for the 36,551 human RefSeq genes (GRCh37p3 assembly) were downloaded from the Ensembl database Genes62 (Wellcome Trust Sanger Institute, UK) using BioMart.[75] From BioMart, we obtained the complete gene coding sequences, together with their associated gene names and description, filtered by RefSeq protein specified with Entrez IDs only. Manning et al.[42] identified nine major classes of human kinases, plus an additional eight atypical classes, altogether a total of 518 genes. Most of these genes occur in single copy; however, nine have paralogs in the human genome. In our analysis, only these nine genes were used, because the paralogs effectively provide repeat measurements and thus statistically more robust estimates of *n*-gram frequencies. Note the nine genes do not correspond to the nine major classes of kinases. The names of these genes, with the number of paralogs for each, are shown in Figure 3. The resulting data set consists of 63 protei*n*-kinase genes taken from the BioMart list. The trigram frequency vector for each paralog group was calculated.

Extracting the corpus of Alu sequences

Genomic locations of *Alu* repeats were obtained from the RepeatMasker track on the UCSC genome browser (www.genome.ucsc.edu).[76] Using the Table Browser service configured by "Variation and repeats" group and family name "*Alu*" as filter value in the "repFamily", we extracted from the human genome draft hg19 a total of 1,091,321 *Alu* repetitive elements with various lengths and their corresponding genomic coordinates. This whole-genome set of *Alu*s was annotated by RepeatMasker (http://repeatmasker.genome.washington.edu) for subfamily association. RepeatMasker identifies 31 *Alu* subfamilies, of which nine were selected for further analysis. These nine were primarily selected for the large number of sequences they contain, to ensure well-defined trigram frequencies. However, they were also selected to include more than one subfamily from each of the three major classes (J, S, Y), and for their frequent usage in other analyses.[45] The four major S subfamilies and the only two known J subfamilies were selected, along with the two mayor Y subfamilies (Ya and Yb). The Yc subfamily was also included because it contains a reasonably large number of sequences and is indisputably younger than Ya and Yb, thus providing an interesting test of the methods ability to discriminate closely related subfamilies. Each of the nine subfamilies contained sequences that varied widely in length due to variable spacer lengths between dimers and variable tail lengths. Such wide variation introduces noise into the *n*-gram frequency vector, so for each subfamily a subset of sequences with relatively homogenous lengths was selected as follows. Sequence lengths in each subfamily were exported to Statistica 7 (Statsoft Inc., Tulsa, OK, USA) and a histogram of lengths was constructed, in each case producing a bimodal length distribution (a known characteristic of *Alu* subfamilies). The bin width for the histogram was 5bp, and only the sequences in the most populated bin were used for subsequent analyses. For example, of the 50,317 sequences in the *Alu*Sq2 subfamily, only the 10,245 sequences of length 295-300bp were selected. The result was a set of 58,122 *Alu* repeats organized into nine sub-families, which were then submitted to the Galaxy server (http://usegalaxy.org) for extraction of the full nucleotide sequences.[77]

Extracting the corpus of promoter sequences

We identified co-expression networks using the gene ontologies examined in our language study (Figure 2). In that study, 9 proteins were grouped into the three diseases Cancer, Neuro and Cardio. The genes encoding those proteins, i.e. HDAC1, HAT1, DNMT1, PSEN2, APP, TUBB3, COX2, MYL2, and MYH7, were used as "seeds" to identify highly correlated promoters in nine co-expression networks. This was accomplished by submitting the gene names to COXPRESdb,[78] which for each seed gene returned the top 300 co-expressed genes. Since co-expression of genes implies shared regulation, the promoter locations (0 to 200 nucleotides upstream from the first exon) of the 300 co-expressed genes for each of the nine networks were obtained from the UCSC genome browser (www.genome.ucsc.edu). The corresponding sets of nucleotide sequences were obtained using the "Fetch Sequences" function of the Galaxy server (http://usegalaxy.org), followed by calculation of the trigram frequency for each promoter network.

n-gram model and frequency vectors

The formal definition of an *n*-gram is:

Given a sequence of n words S = S1S2...Sn over the vocabulary A, and n a positive integer, an *n*-gram of the sequence *S* is any subsequence si...si+*n*-1 of n consecutive words.[79] There are N-n+1 such (not necessarily unique) *n*-grams in *S*. For a vocabulary A with |A| distinct words, there are $|A|^n$ possible unique *n*-grams. In a biological context, *n*-grams can be sequences of n amino acids or nucleotides. For instance, the nucleotide sequence "GGGATC" has two counts of the bigram GG, and one count each of the bigrams GA, AT and TC. To count the *n*-gram frequencies embedded in each language or nucleotide ontology member, a sliding window of length n is run through the sequence from position 1 to n -n +1. The counts are tabulated in the vector Cn for all possible features of length n.

Cn = < cn,1, . . . , cn,*K* > where *K*, the number of all possible features, is 26^nor 4^n and 26 or 4 is the alphabet size, for natural language and DNA sequences respectively. The raw frequency counts are normalized by percent to form a probability distribution vector giving the relative abundance of each *n*-gram. An example of a tri-gram vector is shown in Figure 1. For biological text we used n=2 (bi-grams) and for nucleotides n=3 (tri-grams). Note that the bi-grams in the biological text are counted after all no*n*-alphabetic characters (numbers, spaces and punctuation) are removed.

Construction of n-gram-vectors

The simplest probabilistic models to describe strings like language texts or genomes are low-order Markov models.[80] A zero-order Markov model simply describes the frequencies of each nucleotide (when we consider both strands of genomic DNA, the frequencies of A, T are equal, and so are those of C, G) and the underlying model is a Bernoulli sequence. A first-order Markov model describes the frequencies of individual nucleotides given the nucleotide immediately preceding it; a second-order Markov model describes the frequencies of individual nucleotides given the pair of nucleotides immediately preceding it, and so on. Markov models generate strings that are a poor match to the original genomes, whereas first-order and second-order models do surprisingly well, qualitatively describing the modalities observed in the empirical *n*-gram spectra.[81] It has been previously underappreciated that simple Markov models can generate complex multimodal *n*-gram spectra, probably due to the focus of theoretical research on the distribution of individual *n*-gram counts.

Choosing the Markov order of the model is an important decision. The higher the order, the more accurate the model, up to a point. For example, for English, a second-order model (tri-grams) is more accurate than a first-order (bi-grams) model, which is more accurate than the zero-order (uni-gram) model. But little, if anything, would be gained by using, say, a twentieth-order model rather than a fifteenth-order model. Although much is known about statistical techniques for language recognition in theoretical Markov models, little is known about how well such models and techniques work for real language. Furthermore, higher-order models are more cumbersome than lower-order models, and the space required for vector representations grows exponentially with order.[29,82] For simplicity, we chose the first-order (bi-grams) model of biomedical language, which results in a vector representation of 676 positions. Some readers may wonder: why not simply recognize English by checking if the candidate plain text contains words from an English vocabulary? Indeed, dictionary methods are powerful, when they can be carried out. We chose to not pursue this approach because we are primarily interested in general-purpose statistical methods and because dictionary methods can be analyzed in the context of high-order Markov models. Moreover, it is not possible to use dictionary-based methods with DNA, since a complete DNA dictionary is not available.

In a DNA sequence of four letters, there are 64 possible tri-grams (subsequences of length 3) that can occur, starting from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC, etc. For the purpose of constructing trigram vectors from DNA we introduce a 4 x 4 x 4 cubic matrix with 64 entries, which denote the frequencies of occurrence of all the 64 tri-grams in a DNA sequence. Different *n*-gram distributions (for n=3 to 11) were examined using three different types of DNA corpora (coding (CDS), no*n*-coding (UTR) and randomly generated (RND); Supplementary Figure S5). Zipf-like characteristics of each *n*-gram distribution were examined (data not shown). Tri-grams were chosen to minimize the Mandelbrot fractal effect in the no*n*-Zipfian regimes of frequency population,[83] which in the case of DNA leads to long frequency tails,[81] amplification of nullmers[84] and selection against CpG-containing tri-grams.[85]

Hierarchical cluster analysis

To investigate the degree of dissimilarity (relative distance) in drug/protein, kinase or *Alu* ontologies, the bi-/tri-gram frequency counts were converted to normalized vectors/matrices (by percent) in respect to the total count for all *n*-grams. Hierarchical clustering was performed using the R package pvclust,[86] which assigns statistical significances to clusters using methods developed by Shimodaira[31,32] Various options for distance metric between vectors are provided by pvclust. Here we investigated the two most-used options: Euclidean distance and correlation. Hierarchical clustering is implemented through updating a stored matrix of distances between clusters as each pair of clusters is merged. The distance between clusters can be assessed in various ways, and pvclust includes several options. The two options investigated in this study were complete linkage, which defines the distance between clusters to be the maximum distance between component vectors, and average linkage, which defines the distance between clusters to be the distance between their centroids.[87] The steps in a hierarchical clustering solution that shows the clusters being combined and the values of the distance coefficients at each step are shown by dendrograms. The lengths of the branches in the dendrograms represent distances, and the significance of each cluster is shown on the branches. The package also provides a facility to draw a rectangle around clusters significant at a given threshold, usually 0.95.

Nonnegative matrix factorization (NMF)

The previous step left us with a collection of nw-dimensional frequency-vectors representing the n ontology members in the input list query. The w dimensions represent the w selected *n*-grams. For the 26 letters in the English alphabet used in biological language *w*=676 bigrams, for nucleotides *w* = 64 trigrams. These vectors are arranged as columns of a matrix M of dimensions *w* x n. We use NMF to factor the M matrix into two no*n*-negative lower rank (f) matrices:

M = WH

Where f is the number of factors or ontological features, W is a w x f projection matrix and H is the coefficient matrix of dimension f x n. The column vectors of the W projection matrix are called ontological features, due to the fact that they are collections of related *n*-grams. The columns of H project the original *n*-gram vectors in this new low rank space spanned by the W matrix. To perform the NMF calculations we used the bioNMF web-server application reported in.[88] We used the cophenetic correlation coefficient (ρ, Rho) as well as a clustering heat-map to assess the stability of the factorizations at different ranks (f).[89] The cophenetic correlation coefficient is used as a measure of the robustness of the method in producing stable clusters for a given number of factors (f). Usually the value of f is selected at the point where the magnitude of the cophenetic correlation coefficient shows a significant peak (ρ_{max}). In general, higher values of f will reveal more localized and specific semantic features in the domain. In our case we look for the value f at which the cophenetic correlation coefficient is closest to one (ρ_{max}) because it represents the optimum number of recognized stable features.

Acknowledgement

This work was partially funded by Australian Research Council grant DP1095849, and KG and JWS were supported by the NIH National Technology Center for Networks and Pathways (grant US4RR020843).

Conflict of interest

KG conceived the project, designed the study, obtained and prepared the data, implemented the algorithms, carried out the computational and data analysis including the Statistica and NMF computations, designed and prepared the figures, and drafted and edited the manuscript. SEB contributed substantially to the analysis, interpretation and presentation of all the data, results and figures, ran pvclust and produced the related figures, and made extensive contributions to all aspects of the manuscript. JWS participated in data and figure discussions and interpretations, and conceived the extension of the study into gene regulation and promoter networks. JMK instigated the pvclust analyses and contributed to the interpretation of those results, guided revision of statistical aspects of the paper, and extensively redrafted the manuscript.

References

1. Richards RJ. *Darwin and the emergence of evolutionary theories of mind and behavior*. USA: University of Chicago Press; 1989. p. 718.
2. Darwin C. *The expression of the emotions in man and animals.* 1st ed. John Murray, London; 1872. 374p.
3. Mantegna RN, Buldyrev SV, Goldberger AL, et al. Linguistic features of noncoding DNA sequences. *Phys Rev Lett.* 1994;73(23):3169–3172.
4. Pesole G, Attimonelli M, Saccone C. Linguistic approaches to the analysis of sequence information. *Trends Biotechnol.* 1994;12(10):401–408.
5. Popov O, Segal DM, Trifonov EN. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems*. 1996;38(1):65–74.
6. Botstein D, Cherry JM. Molecular linguistics: extracting information from gene and protein sequences. *Proc Natl Acad Sci U S A*. 1997;94(11):5506–5507.
7. Brendel V, Beckmann JS, Trifonov EN. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *J Biomol Struct Dyn*. 1986;4(1):11–21.
8. Israeloff NE, Kagalenko M, Chan K. Can Zipf distinguish language from noise in noncoding DNA? *Phys Rev Lett.* 1996;76(11):1976–1976.
9. Dassow J, Mitrana V. Evolutionary grammars: A grammatical model for genome evolution. *Bioinformatics*. 1997;1278:199–209.
10. Zipf GK. *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge: Addison-Wesley Press; 1949. p. 573.
11. Gamow G, Ycas M. Statistical correlation of protein and ribonucleic acid composition. *Proc Natl Acad Sci U S A*. 1955;41(12):1011–1019.
12. Searls DB. The roots of bioinformatics. *PLoS Comput Biol.* 2010;6(6):e1000809.
13. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*. 1998;8(3):163–167.
14. Vinga S, Almeida J. Alignment-free sequence comparison-a review. *Bioinformatics*. 2003;19(4):513–523.
15. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science.* 2008;319(5862):473–476.
16. Kemena C, Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*. 2009;25(19):2455–2465.
17. Randic M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. *Chemical Physics Letters*. 2006;419(4–6):528–532.
18. Gu X, Zhang H. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol*. 2004;21(7):1401–1408.
19. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. *Nat Genet.* 1999;21(1): 108–110.
20. Zhang H, Zhong Y, Hao B, et al. A simple method for phylogenomic inference using the information of gene content of genomes. *Gene*. 2009;441(1–2):163–168.
21. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*. 2003;19(16):2122–2130.
22. Yang L, Zhang X, Wang T. The Burrows-Wheeler similarity distribution between biological sequences based on Burrows-Wheeler transform. *J Theor Biol*. 2010;262(4):742–749.
23. Zhang S, Wang T. A complexity-based method to compare RNA secondary structures and its application. *J Biomol Struct Dyn*. 2010;28(2):247–258.
24. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 1986;83(14):5155–5159.
25. Karlin S, Ladunga I. Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A*. 1994;91(26):12832–12836.
26. Sims GE, Jun SR, Wu GA, et al. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A*. 2009;106(8):2677–2682.
27. Hackenberg M, Carpena P, Bernaola-Galvan P, et al. WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms Mol Biol*. 2011;6:2.
28. Chomsky N. Syntactic structures. Berlin, New York: Mouton de Gruyter; 2002. p. 117.
29. Pinello L, Lo Bosco G, Yuan GC. Applications of alignment-free methods in epigenomics. *Brief Bioinform*. 2013.
30. Kotamarti RM, Hahsler M, Raiford D, et al. Analyzing taxonomic classification using extensible Markov models. *Bioinformatics*. 2010;26(18):2235–2241.
31. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 2002;51(3):492–508.
32. Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of statistics*. 2004;32(6):2616–2641.
33. Brunet JP, Tamayo P, Golub TR, et al. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164–4169.

34. Lee DD, Seung HS. Learning the parts of objects by no*n*-negative matrix factorization. *Nature*. 1999;401(6755):788–791.

35. Kim H, Park H. Sparse no*n*-negative matrix factorizations via alternating no*n*-negativity-constrained least squares for microarray data analysis. *Bioinformatics*. 2007;23(15):1495–1502.

36. Kim H, Park H, Drake BL. Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. *BMC Bioinformatics*. 2007;8(Suppl 9):S6.

37. Heinrich KE, Berry MW, Homayouni R. Gene tree labeling using nonnegative matrix factorization on biomedical literature. *Comput Intell Neurosci.* 2008.

38. Chagoyen M, Carmona-Saez P, Shatkay H, et al. Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*. 2006;7:41.

39. Ulrey CL, Liu L, Andrews LG, et al. The impact of metabolism on DNA methylation. *Hum Mol Genet.* 2005;14 Spec No 1:R139–147.

40. Grassot J, Gouy M, Perriere G, et al. Origin and molecular evolution of receptor tyrosine kinases with immunoglobuli*n*-like domains. *Mol Biol Evol*. 2006;23(6):1232–1241.

41. D'Aniello S, Irimia M, Maeso I, et al. Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus. *Mol Biol Evol*. 2008;25(9):1841–1854.

42. Manning G, Whyte DB, Martinez R, et al. The protein kinase complement of the human genome. *Science*. 2002;298(5600):1912–1934.

43. Martin J, Anamika K, Srinivasan N. Classification of protein kinases on the basis of both kinase and no*n*-kinase regions. *PLoS One.* 2010;5(9):e12460.

44. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by populatio*n*-scale genome sequencing. *Nature.* 2011;470(7332):59–65.

45. Batzer MA, Deininger PL, Hellman*n*-Blumberg U, et al. Standardized nomenclature for *Alu* repeats. *J Mol Evol*. 1996;42(1):3–6.

46. Kapitonov V, Jurka J. The age of *Alu* subfamilies. *J Mol Evol*. 1996;42(1):59–65.

47. Jayadev S, Leverenz JB, Steinbart E, et al. Alzheimer's disease phenotypes and genotypes associated with mutations in presenilin 2. *Brain*. 2010;133(Pt 4):1143–1154.

48. Funada A, Konno T, Fujino N, et al. Impact of Reni*n*-Angiotensin System Polymorphisms on Development of Systolic Dysfunction in Hypertrophic Cardiomyopathy -Evidence From a Study of Genotyped Patients. *Circulation Journal*. 2010;74:2674–2680.

49. Li Y, Wu G, Tang Q, et al. Slow cardiac myosin regulatory light chain 2 (MYL2) was dow*n*-expressed in chronic heart failure patients. *Clin Cardiol.* 2011;34(1):30–34.

50. Osborne JD, Flatow J, Holko M, et al. Annotating the human genome with disease ontology. *BMC Genomics*. 2009;10(Suppl 1):S6.

51. Flam F. Hints of a Language in Junk DNA. *Science*. 1994;266(5189):1320.

52. Doerfler W. In search of more complex genetic-codes -Can linguistics be a guide. *Medical Hypotheses*. 1982;9(6):563–579.

53. Bonhoeffer S, Herz AVM, Boerlijst MC, et al. Explaining ''linguistic features'' of noncoding DNA. *Science.* 1996;271(5245):14–15.

54. Shannon CE. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948;27:379–656.

55. Shulman MJ, Steinberg CM, Westmoreland N. The coding function of nucleotide-sequences can be discerned by statistical-analysis. *J Theor Biol.* 1981;88(3):409–420.

56. Claverie JM, Bougueleret L. Heuristic informational analysis of sequences. *Nucleic Acids Res.* 1986;14(1):179–196.

57. Karlin S, Cardon LR. Computational DNA-sequence analysis. *Annu Rev Microbiol*. 1994;48:619–654.

58. Ciccarelli FD, Doerks T, von Mering C, et al. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science.* 2006;311(5765):1283–1287.

59. Konopka AK, Martindale C. Noncoding DNA, Zipf's law, and language. *Science.* 1995;268(5212):789.

60. Chatzidimitriou-Dreismann CA, Streffer RMF, Larhammar D. Lack of biological significance in the 'linguistic features' of noncoding DNA--a quantitative analysis. *Nucleic Acids Res*. 1996;24(9):1676–1681.

61. Breslauer KJ, Frank R, Blocker H, et al. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A.* 1986;83(11):3746–3750.

62. Hunter CA. Sequence-dependent DNA structure. The role of base stacking interactions. *J Mol Biol.* 1993;230(3):1025–1054.

63. Fernald GH, Capriotti E, Daneshjou R, et al. Bioinformatics challenges for personalized medicine. *Bioinformatics*. 2011;27(13):1741–1748.

64. Rizzo JM, Buck MJ. Key principles and clinical applications of "Next-Generation" DNA sequencing. *Cancer Prev Res*. 2012;5(7):887–900.

65. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.

66. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17(4):540–552.

67. Kolmogorov A. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*. 1968;14(5):662–664.

68. Tsuruoka Y, Tsujii J, Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics.* 2008;24(21):2559–2560.

69. Perez-Iratxeta C, Bork P, Andrade MA. Exploring MEDLINE abstracts with XplorMed. *Drugs Today (Barc)*. 2002;38(6):381–389.

70. Lin SM, McConnell P, Johnson KF, et al. MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics.* 2004;20(18):3659–3661.

71. Maier H, Dohr S, Grote K, et al. LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Res.* 2005;33:W779–782.

72. Jelier R, Schuemie MJ, Veldhoven A, et al. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol*. 2008;9(6):R96.

73. Trieschnigg D, Pezik P, Lee V, et al. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics.* 2009;25(11):1412–1418.

74. Zhu S, Zeng J, Mamitsuka H. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. *Bioinformatics.* 2009;25(15):1944–1951.

75. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005;21(16):3439–3440.

76. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996–1006.

77. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010;11(8):R86.

78. Obayashi T, Hayashi S, Shibaoka M, et al. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*. 2008;36:D77–D82.

79. Manning CD, Schutze H. *Foundations of statistical natural language processing*. Cambridge: MIT Press; 1999. p. 680.

80. Norris JR. *Markov Chains*. UK: Cambridge University Press; 1998.

81. Chor B, Horn D, Goldman N, et al. Genomic DNA k-mer spectra: models and modalities. *Genome Biol*. 2009;10(10):R108.

82. Martin S, Liermann J, Ney H. Algorithms for bigram and trigram word clustering. *Speech Communication*. 1998;24:19–37.

83. Montemurro MA. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*. 2001;300:567–578.

84. Garcia SP, Pinho AJ, Rodrigues JM, et al. Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS One*. 2001;6(1):e16065.

85. Acquisti C, Poste G, Curtiss D, et al. Nullomers: really a matter of natural selection? *PLoS One*. 2007;2(10):e1022.

86. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* 2006;22(12):1540–1542.

87. Anderberg MR. *Cluster analysis for applications*. New York: Academic Press; 1973. 359p.

88. Mejia-Roa E, Carmona-Saez P, Nogales R, et al. bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Res*. 2008;36:W523–528.

89. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol.* 2008;4(7):e1000029.

Citation: Gramatikoff K, Boyd SE, Smith JW, et al. Genomic taxonomy boost by lexical clustering. *J Investig Genomics.* 2014;1(1):13–25.
DOI: 10.15406/jig.2014.01.00004