Computational and Structural Biotechnology Journal 11 (2014) 123–130

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

RNA-Seq Data: A Complexity Journey

Enrico Capobianco *

Center for Computational Science, University of Miami, Miami, FL, USA
Laboratory of Integrative Systems Medicine, IFC-CNR, Pisa, Italy

ARTICLE INFO

Available online 17 September 2014

Keywords:
Transcriptome profiling
RNA-Seq
Complexity
Inverse problems
Networks

ABSTRACT

A paragraph from the highlights of "*Transcriptomics: Throwing light on dark matter*" by L. Flintoft (Nature Reviews Genetics 11, 455, 2010), says: "Reports over the past few years of extensive transcription throughout eukaryotic genomes have led to considerable excitement. However, doubts have been raised about the methods that have detected this pervasive transcription and about how much of it is functional." Since the appearance of the ENCODE project and due to follow-up work, a shift from the pervasive transcription observed from *RNA-Seq* data to its functional validation is gradually occurring. However, much less attention has been turned to the problem of deciphering the complexity of transcriptome data, which determines uncertainty with regard to identification, quantification and differential expression of genes and non-coding RNAs. The aim of this mini-review is to emphasize transcriptome-related problems of direct and inverse nature for which novel inference approaches are needed.

Contents

1. Background

Next Generation Sequencing (NGS) and *Big Data* are companions in the fascinating life science era. The great impacts in biomedicine which are expected from their clever exploitation are stimulating exciting research at all latitudes. Given the fact that a growing number and variety of data-intensive NGS applications are also expected, and that technologies are constantly subject to refinements, the estimation of data-robust information bounds establishing a reference or benchmark for the quality, reliability and significance of the results is destined to remain a purely theoretical task. In the meantime, debating the current knowledge gaps can be also useful, and central to our review is the observation that the entire RNA cellular component is comprised by the transcriptome [1], implying that the estimation of the expressed transcripts will determine the regulation networks underlying key phenotypes, in particular in relation with disease.

* Center for Computational Science, University of Miami, Miami, FL, USA.
E-mail address: ecapobianco@med.miami.edu.

http://dx.doi.org/10.1016/j.csbj.2014.09.004

Processing and analyzing transcriptomes, especially human ones, require the assimilation of large volumes of data into computational pipelines designed for specific inference tasks, i.e., identification, quantification, differential expression, profiling, annotation, prediction etc. However, when looking at the targets of such tasks, even for the transcripts assumed to be known, e.g. the protein coding genes, there is not yet clear consensus about what could be defined a representative number of them. The knowledge gap further extends when structural complexity (small-large non-coding RNAs, novel transcripts from non-annotated genes, splicing isoforms etc.) comes into the picture, and redundancy and noise need to be considered.

Several complexities arise from transcriptome measurements, and these transfer over the data. *RNA-Seq* [2–4] deliver transcriptome snapshots by estimating the copy number of the transcripts in samples, and the results allow for accurate digital gene expression measurements and prediction of novel transcripts. Two main problems are worth mentioning:

1. *Read mapping uncertainty* [5]. In general, reconstructing full-length transcripts requires an assembly phase, except for the small RNAs that are often shorter than the already short sequence reads obtained from the common platforms. Therefore, ambiguity occurs because of a limited resolution, and from the fact that transcript variants originated from the same gene may share exons.
2. *Coverage* [6], which varies across transcripts as a consequence of the fact that transcripts are expressed at different levels and have different lengths (among other factors producing biases, like GC content, for instance). Therefore, transcripts that are consistent with known isoforms may remain incompletely assessed due to limited coverage, and lowly or broadly expressed genes are less supported by *RNA-Seq* compared to abundant transcripts, which are instead fully assembled.

Other known problems are sequencing non-uniformity, estimation of novel isoforms (alternatively spliced transcripts), and quantification of expression levels. While it appears likely that novel complexities will emerge from data structures generated by newly designed experiments, it is important to consider the data available in public repositories, as they are becoming natural target of scientific reuse. This data multitude requires integration strategies to deal with heterogeneous sources and categorized entities, thus suggesting the need of identifying specific features in variables and parameters which should shape the spectrum of inference tools.

A few questions can be formulated with regard to establishing a rationale behind the choice of an inference approach for *RNA-Seq* data: *a) Given a certain problem complexity, what is the best possible approach in terms of reproducibility of solutions? b) What conditions should lead to model-based versus model-free (data-driven) methods? c) How to ensure that statistical estimates are reasonably accurate?*

It is generally known that solving an *inverse problem* entails determining unknown *causes* based on the observation of their *effects*, unlike for a *direct problem* in which the solution involves finding effects based on a complete description of their causes. One of the least noticed aspect in 'omics' applications is that many problems have an inverse nature [7]. For these, ad hoc statistical inference may solve the convolutions between complex multi-parameter variables, and new strategies may involve the multilevel power of networks, as explained below.

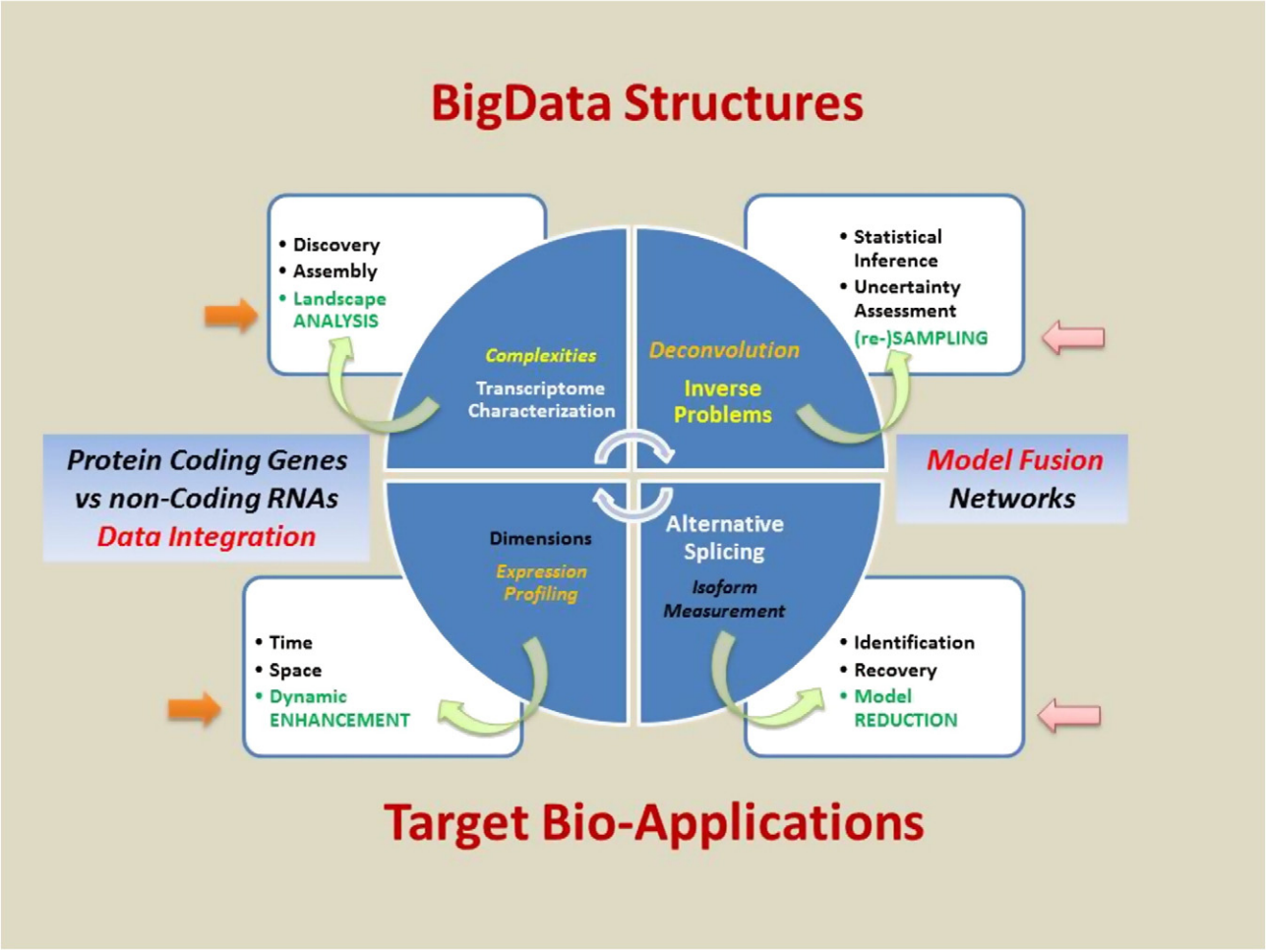

Fig. 1. Concepts describing complexity, and corresponding architecture.

2. Reading through complexity

Fig. 1 provides the reference for the proposed analysis on complexity underlying *RNA-Seq* data, and suggests a series of problems, each specified in detail. The informative layers are three:

i) Top layer: *Big Data* structures;
ii) Middle layer: a Cartesian coordinate system with a view of translational topics; and
iii) Bottom layer: target bio-applications.

In particular, a system's view would embed these three components and establish for the top layer the function of data generation, for the middle layer the role of the methodological core whose complex design is dissected into quadrants linked to tasks, actions and applications, and for the bottom layer the systems outcome.

Problem #1. Sequencing depth. Regulation of such factor at the experimental level is known to influence the depth of discovery achievable from the data, thus determining the impact of computational methods.

The first quadrant includes terms like *inverse problem* and *deconvolution* (see also Box 1), which provide methods for model identification, estimation and testing. Assessment of uncertainty for the measured values is also reported. In particular, *re-sampling* deals with testing data structures for the presence of real and significant features. Statistical techniques offer many solutions. One strategy involves building *null models* as a benchmark for the validation of results. Sampling and re-sampling (SRS) techniques point to well-known solutions such as jackknife, bootstrap, permutations and cross-validation. In some cases, when the parametric assumptions are in doubt, the risk of mis-specification is high, and systematic analytical efforts are required to compute standard errors of the estimates.

Problem #2. Transcript structures include protein-coding genes and non-coding RNA biotypes that need to be cross-examined to annotate as many entities as possible that connect genotype to phenotype of a cell. The examination of transcripts arising from intergenic regions (i.e. lincRNA) should help to elucidate functional relevance by generating testable hypothesis subject to further validations.

The second quadrant includes *complexities*. Among their determinants, *data dimensionality* (low-high) and *scale* (poor–rich), together with *variables' heterogeneity* with both *functional relationships* (linear–nonlinear) and *parameter settings* (finite–infinite), are among the most important entities. The transcriptome characterization involves both the *discovery* of biotypes, and an *assembly* phase with accurate and quantitative mapping of billions of reads to small/large reference genomes, aiming at the simultaneous reconstruction of all full-length mRNA transcripts from the available short/long reads.

Problem #3. The multidimensionality of transcriptome profiling is a key factor. While estimating dynamically the expression profiles, mapping them to gene and protein networks can help driving inference on yet unexplored RNA functional regulation of the transcriptome.

The third quadrant addresses *dimensions* that are relevant for *expression profiling*; these are *time and space*. Data structures embedding both dimensions require particular experimental designs, which are rarely accessible from public resources. Treating data in the presence of both dimensions requires an adaptation of current computational algorithms: see for instance the proposals of [8,9] in which complex biological variability makes both statistical multifactoriality and robustness central to the analysis. In large part, innovative solutions are currently unavailable to address a *dynamic enhancement* of expression profiles toward spatiotemporal extensions. Among the possible reasons, one refers to the data structure, especially with reference to time course profiles embedding dependencies that cannot be treated based on data replication strategies or assuming independence that are of common use in simple condition-specific dynamics.

Problem #4. Evidence of identified isoforms represents a knowledge basis on variation. Model reduction techniques need to be applied to the isoform space to allow recovery and prediction.

The fourth quadrant includes *alternative splicing* (AS), and addresses the related problem of measuring and modeling isoforms. AS (see [10], and the inserted references n. 1–15 for functional role and references n. 16–25 for involvement in disease, and [11] for variability in human populations) is a common mechanism of gene regulation in higher eukaryotes, which occurs in over 90% of multi-exon genes in the human genome and is regulated by complex interactions between cis-acting splicing elements and trans-acting factors. Many splicing regulators have tissue-specific expression patterns, resulting in widespread differences in AS patterns across different tissues. In connection with Problem #3, the consideration of AS spatiotemporal dynamics may help isoform identification and recovery problems [12–16]. In general, depending on the assumptions about the isoforms, a *model reduction* strategy is considered relevant, i.e. a subset of measurable isoforms spanning or approximating the included isoforms (also the unknown ones). In most cases the isoform space is a hidden information layer destined to remain ignored, and thus what is empirically found from the sequence processing is redundancy difficult to handle or even interpret.

Box 1

Dimensional model reduction

Methods such as compressed sensing/sampling (CS), principal/independent/sparse component analysis, latent factor analysis represent well-known dimensionality reduction techniques adopted in statistical inference and machine learning. CS modeling simplifies the 'reduction' task compared to other probabilistic alternatives which are usually based on maximum likelihood estimation, bringing thus the risk of inconsistent results in dependence on the nature of the data.
Model mis-specification is a possibility whenever distributional assumptions (say, Gaussianity) do not hold, and in such cases different likelihood functions can be associated with the data (e.g. pseudo-likelihood, profile likelihood, and penalized likelihood) thus switching from model-based to data-driven approaches. However, some problems are naturally 'inverse', and call for regularized solutions to correct for their 'ill-posedness', e.g. their solution is unstable under data perturbation. Numerical methods thus can cope with such limitations, and CS optimization is suitable for such task.

Deconvolution

Signal deconvolution methods represent a possible strategy to address inverse problems, and are employed to recover an acceptable signal-to-noise ratio in situations in which this is too low due to convolution of signal with response and other systems interferences. These methods call for projective approaches in statistical applications, usually relying on non-parametric inference solutions, and also involve a sequence of algorithmic steps such as filtering, smoothing and prediction when signal processing problems are under investigation.

The following hypothesis-driven research actions are the most promising ones:

a. *Analyze human transcriptome complexities by characterizing multidimensional expression profiles, pathway landscapes and wide-spectrum transcript structures.*
b. *Provide better transcript annotations, possibly extending them with new ones;*
c. *Select inference approaches for inverse problems targeted to both isoform deconvolution (transcript quantification) and isoform prediction;*
d. *Elucidate role and function of isoforms to provide further insights into the unknown transcriptome functionalization.*

3. Multiscale view

Transcriptome-related complexities are difficult to decipher, analyze and predict due to their co-dependence over at least three scales: *data, methods and applications*. Statistical complexity [17] refers to the presence of non-trivial hidden patterns in the system's dynamics, reflecting an information-disequilibrium interplay. Consequently, we should first account for statistical fluctuations in a system observed through finite samples and then, from the interdependence between complexity and distance, address the relationships between dynamic patterns and fluctuations, as the system's outcomes may occur with empirical (observed) frequencies (probabilities) which can be different from the expected (theoretical) ones.

In general, given an experimental data generation process, any replicated data should aim to produce measurements with controlled variation, i.e. bounded by the average fluctuation. However, the assessment of such variation bounds to discriminate between transcriptional entities and noise remains highly uncertain. Usually, algorithmic and data intensive *RNA-Seq* pipelines of both probabilistic and heuristic specifications are employed agnostically, following two main directions:

1) *Customization of model solutions to experimentally generated data (narrowing down the problem)*. Results which are difficult to generalize involve the analysis of degree of overfitting possibly present in the data. Examples of this form of bias are provided by typical parametric models, which are based on model-specifying assumptions.

2) *Adaptation to data features (relaxing the problem from constraints)*. Often, the complexity cannot be properly handled, because the problem is not sufficiently regularized. Examples are provided by solutions arising from non-parametric models, which are simply data-driven.

The consideration of the multiscale nature of expression profiling is often overlooked. Regarding time-related dependencies, a recent advance is worth mentioning [18]. Spatiotemporal dynamics are not directly covered in *RNA-Seq* pipelines, rising doubts about how to algorithmically handle with numerical precision and statistical accuracy the time-course information and/or the spatial heterogeneity in many data-intensive applications. However, systems-level research targeted to complex biological systems and their pathological alterations is depending on spatiotemporal analysis.

For instance, recent advances in single cell technologies are contributing to novel transformative approaches and challenges in computational biology (two recent examples are [19,20]). This is due to the new possibilities of improving the task of monitoring biological systems at multiple temporal, spatial and molecular resolutions at the single cell level. Single cell profiling approaches allow the screening of transcript amounts referred to many different markers, enabling the evaluation of whole transcriptomes and genomes for single cells. Consequently, information about complex phenotypes investigated in heterogeneous cell populations will be possible; an example is cancer and its well-known altered hierarchy of cell sub-populations for which future perspectives are to find novel therapeutic targets, to discover new functionally relevant cell types, and to infer cell-to-cell variability phenomena in a variety of cell populations.

Problems involving the detection of unknown isoforms and the estimation of their expression values, usually refer to a given isoform space. Finding a solution path in complex high-dimensional data spaces requires model reduction approaches to handle the dimensionality (see Box 1). In the context of isoform modeling, solutions may be adapted to a rich class of analysis methods involving projections in new coordinate spaces from which to assess the systems dependencies. In particular, the domain of *compressed sensing/sampling* (CS) techniques [21,22] is yet to be explored, together with other methods dealing with *principal and independent components*, or with *sparse or latent factors*.

The rationale for such applications is offered by the advantage of exploiting the inherent partial knowledge of the isoform vector attached to each gene, i.e. the fact that M of the N given values, with $M < N$, are quantified and the rest are not (the unknown part of the vector). The common practice of considering a parsimonious set of isoforms based on the expression values is not exempt from risks, due to the fact that other informative isoform features (in relation to structure, localization, etc.) would be excluded by a selection criterion based just on expression values. Such features could then be included in the *sensing matrix* for determining the measurements of the isoforms. CS modeling would simplify the reduction problem compared to probabilistic alternatives based on maximum likelihood types of estimation, which may result inconsistent with the nature of the data.

However, the inverse nature of the problem suggests that regularized solutions might be needed. This aspect can be explored within a CS optimization context. In particular, an isoform stochastic context addresses equivalently a signal recovery problem in which the observed variables are modeled through latent or hidden components. When the latter exist, they need to be unambiguously identified and estimated. The other way to look at this class of problems is to consider data decompositions, and in our case the *RNA-Seq* landscape would be the candidate for such decomposition aimed at extracting underlying signatures. Assessing the impact of AS on the expression profiles could be a goal for a detection algorithm, but other signatures could be targeted too, including transcriptional noise whose role remains ambiguously present when studying non-coding RNA structure characterizations.

4. Transcriptome landscape analysis and structure recovery

The complex layers emerging from analyses at transcript scale are substantially convoluted, preventing from a clear identification of separate effects. The overall complexity translates in part into pervasive transcription, functionalization, and integrative regulation. For instance, pervasive transcription [23,24] involves complexity factors such as: i) Regions characterized by long ncRNA [25], with specific interactions and functions; ii) LincRNA [26–30], and for applications see [31–33] which are currently not covered by annotation; iii) Transcriptional noise; and iv) Fragments of known pre-mRNA. A full elucidation of such aspects is not trivial.

One problem is that lowly/broadly expressed transcripts are weakly/incompletely supported by *RNA-Seq*; while the abundant transcripts are represented by many reads, the rare transcripts are represented by only a few reads. For instance, while in general lincRNAs show less abundance than protein-coding genes, are less expressed and have a tissue/cell-specific expression pattern, these RNAs are multi-exon transcripts mapping to intergenic regions, in which most of transcribed mRNAs do not encode proteins.

Overall, many *RNA-Seq* experiments have shown read density more than 100-fold higher in exons than in introns or intergenic regions. Despite the biological role of most of the intergenic transcripts still remains poorly understood or unknown, many of them possess functional roles. Indeed, the genomic loci of many lincRNA lie close to neighborhood genes, presenting smaller size and shorter transcripts. Thus, it has been proposed a regulatory role of these transcripts on proximal genes, and a potential role on AS of mRNA isoforms.

In addition, many of the lincRNA loci transcribe different AS isoforms, in agreement with the current knowledge that about 90% of genes generate alternative mRNA isoforms. mRNA complexity refers to the extent by which alternative RNA isoforms contribute to functional diversity. It is known that a small number of RNA regulatory proteins may produce a vast diversity of biological outcomes. These considerations are confirmed by studies such as ENCODE, showing that known isoforms account for about 80% of *RNA-Seq* fragments due to highly expressed genes that are involved, while a residual 11% of the fragments map to novel isoforms of known genes.

5. What's next?

Complementary methods are needed to discriminate between functional low-abundance transcripts and both transcriptional noise and process artifacts in regions of transcription detected by *RNA-Seq*. Such discrimination would aim to produce a reliable testable hypothesis subject to validation. Signal deconvolution deals with ill-posed problems [7] derived from the convolution of splice variants through read mapping. In order to balance any insufficient sampling of the transcriptome (i.e. low expressions preventing from detecting structures), the consideration of how to transform low expression data to provide improved transcript re-capture may be worthy. This step would involve change-of-coordinate (projective) approaches or re-scaled (transformed) models and several filtering, smoothing and prediction algorithms belonging to the realm of non-parametric inference, i.e. data-driven rather than based on probabilistic assumptions.

A field of natural application would especially be dark matter [34]. This term indicates sequences of unidentified type and ill-determined functions, which can contain both coding and ncRNA as long as the functions remain unclear, but also refer to the transcription from intergenic regions of annotated genes. An open question is whether the ncRNAs are biologically relevant or not. If suspected to be irrelevant, why their functions appear persistent? If so, how their expression patterns can be distinguished from the signature of transcriptional noise? One of the hypotheses underlying such novel transcripts is that they are a by-product, rather than an independent functional unit. Dark matter could arise from extended or complex transcription of known genes, may promote the transcription of neighboring protein-coding genes, and arise from regions predicted to contain open chromatin, suggesting possible regulatory roles for ncRNAs regulating epigenetic memory through modifications to DNA and chromatin structure [35–37].

Read mapping of *RNA-Seq* to intergenic regions displays correlation with proximal genes or annotation with novel exons. Intergenic regions present usually a mix of potentially coding (extended transcripts) and separated non-coding transcripts. Complications arise because the true number and level of different transcript isoforms are not usually known, and transcription activity varies across the genome. Transcription is considered an efficient regulatory process in cells, organisms and tissues which controls the complex form of gene expression. Specific events such as pausing (and backtracking) affect transcription, and the heterogeneity in transcription rate makes it not a continuous process, but a process subject to interruptions with negative effects on transcription. In this mini-review, the goal is to assess the power of data-driven inference in uncovering transcriptome features. From a computational side, read mapping uncertainty involves the problem that reads do not span entire transcripts, thus the transcripts from which they are derived are not always uniquely determined and many reads align to multiple transcripts. Consequently, an identifiability problem occurs, which in the context of *RNA-Seq* data translates into different parameters (relative transcript abundances) generating different probability distributions on the read counts, and in turn affects the transcript assembly task whose relative abundance estimation is limited by the transcriptome incompleteness. If each target genetic feature is considered a particular entity for which the population size has to be estimated, the problem of sequencing sampling can be seen as random sampling of each entity aimed to estimate the relative abundances in the corresponding population. SRS statistical techniques can offer a solution to validate the detected structures against null models, and thus test first their statistical, and then their biological significance. The paradigm shift thus redefines the sampling process of *RNA-Seq*, from endogenous sampling (biological and technological types) to post-processing re-sampling (statistical type).

6. Alternative splicing & isoform modeling

AS is a process by which a single DNA sequence can be transcribed in multiple mRNAs. When this process occurs in protein-coding genes it enhances the transcriptome (and in turn the proteome). Despite a majority of multi-exon genes undergoing AS in a tissue-specific way, the fraction of functional versus spurious splicing remains undetermined. At the protein level, AS isoforms may have different patterns. It is known in part how to use *RNA-Seq* data to infer the existence of novel isoforms in known transcribed regions. The accurate mapping of reads that span splice junctions is a crucial step in *RNA-Seq* data analysis. Reads are usually short and can map to multiple isoforms at the same time. Thus, genes with multiple isoforms complicate the task of determining which (known or unknown) isoform produced each read (isoforms are sampled non-uniformly, i.e. with probability proportional to their length). Consequently, for some genes the isoform expressions are non-identifiable and cannot therefore be estimated separately. The following are challenging problems:

1. *Most generated splice variants have to be evaluated, starting from one aspect — are they constitutive or not?*
2. *For the genes expressing multiple splice variants there is uncertainty about the predominant one in any tissue or cell type;*
3. *The dynamic nature of AS implies that fluctuations characterize isoforms at different timescales, affecting model reduction and selection.*

It is expected for only a small number of factors to be involved in any AS event, thus implying an overall control exerted by relatively few regulators. In turn this would lead to a simplification of the problem of identifying isoform diversity and providing recovery from various expression levels. However, there's no universally recognized best computational method for inferring isoforms from short reads.

7. What models to use?

Part of the complexity derives from the fact that different isoforms can generate common reads that collapse over the multiple transcripts producing them. Approaches centered on *inverse problems* to estimate isoforms are a potential direction to follow. Inverse problems call for models contemplating the presence of both observable and latent variables; these latter unknown components must be identified and estimated, thus permitting a successful recovery. Some of the key properties required are related to data (sparsity), models (compressibility) and methods (efficiency). As previously stated, CS allows to efficiently reconstruct an unknown vector of N entries from only $k < N$ measurements. Therefore, it leverages on the concept of sparsity. Translated into the isoform context, CS would consider a vector X of isoforms associated with a gene of unknown length N. The Y measurements of length M ($M < N$), are associated with entries whose values depend on observed isoform features, such as measured expression. M is the sensing isoform matrix carrying information through measurable features.

The number of reads can be proxy of some entities at a given locus, for instance the number of isoforms of a certain gene. Such number would provide an estimate of the number of isoforms, i.e. the length N of the vector. Outside a coding region, and proportionally to the overall abundance, we might observe the presence of a number of reads that might be targeting the same gene, in case the latter lacks sufficient approximation power (coverage), or otherwise indicates ncRNA transcriptional activity. The goal should be to infer isoforms when they are

sufficiently expressed. Thus, a threshold should act as a sparsity control mechanism, without penalizing low-expressed but functionally relevant transcripts. Then, it would be worth considering the attempt of heading a minimal isoform set that is able to explain the read data by multiple features, while using dimensionality reduction algorithms to make it a feasible goal.

8. Expression profiling (multidimensionality and dynamic enhancement)

Mixtures of hidden isoforms form a dynamic layer of the transcriptome, which combined with changes in expressions, can influence functional analysis. Gene expression and regulation are intrinsically stochastic and noisy processes responsible for the stochastic variation in the transcriptome. Processing *RNA-Seq* data in dependence of conditions referred to different time points, requires consideration of temporal dependence to ensure that the most likely estimable parameters maximize the probability of observing the experimental data (e.g. the maximum likelihood estimates). As more variables and parameters increase the complexity of the model, penalties should be used to discount the complexity by inducing regularization.

9. How to deal with dynamics?

Biological systems have been investigated also with respect to the reconstruction of the phase space of the generating dynamical process [38,39]. A dynamical system is assumed to perform a trajectory in a state-space spanned by the gene expression levels (or mRNA concentrations) representing the state space variables. This way, a time-course profile could be considered dynamically enhanced when the time-ordered levels of gene expression detected at each phase are concatenated to form a temporal record of measurable quantities. Several projective and regularized methods could be applied to separate signal from noise, to identify components, and to estimate the correlation. Phenotypes match some regions of the gene expression signatures in the state-space; the signature is a specific feature, and the multidimensional landscape includes each phenotype representing a point or an example. When projective decomposition techniques are applied, the identified components should reflect dysregulated pathways, due to the perturbations, and persistent versus transient profiles from the gene sets enriching for the pathways. When n states are observed, with n big, an ensemble is obtained thus allowing for consideration of steady dynamics. Before steady-state dynamics, transient dynamics can be observed and such non-equilibrium conditions require statistical methods to infer the important mechanisms behind such dynamics.

10. Differential network analysis

Differential network biology [40] proposes novel approaches based on the comparison between topological configurations and modular structures in health versus disease states, or given two different disease states or perturbations, often supported by expression, genetic, and clinical information. In particular, the variation may reflect the distinctiveness in molecular signatures of gene expression and protein translation arising from different combinations of genetic mutations. Once such signatures are assigned to network nodes, topological and biological features may elucidate interactions and causal relationships. Particularly when molecular complexes or signaling pathways are considered, the identification of cancer-related hubs and interface proteins (in topological terms), or upstream signals and downstream targets (in pathway localization terms), may involve differential connectivity or gene co-expression (possibly co-regulation) patterns that lead to a sub-network or pathway-centric marker classification [41–43]. Targeting altered signaling networks can suggest novel therapeutic strategies classified within the field of network medicine [44]. For instance, tumor progression involves signaling network robust rewiring for the transmission of phenotypic alterations, which implies that module-coordination occurs as a system's level response [45].

11. What's at play?

Spatiotemporal dynamics are rarely represented within network maps, usually replaced by averages taken over conditions or time points, despite the importance of revealing their inherent potential to represent aggregates of many entities in simultaneous relationships. Once the dynamic data sources are generated, say through gene expression profiles, active network components may be identified by direct extension (with genes) or simply by mapping (with proteins). Repeating in parallel this operation at each given time, say, would generate transcript profiling coupled with network configuration profiling, where in the latter comparisons between components; i.e. active versus unaffected ones, can be performed at each step. These components may represent relevant summaries of biological activities, such as pathways, obtained through the system's dissection by projective techniques (i.e. decorrelated, independent, etc.) whose profile can be monitored step by step. In general, changes in network configurations may refer to the structure of modularity, in terms of both module composition and inter-module connectivity patterns.

Different types of networks are known, for instance: i) Gene regulatory networks; ii) Protein interactome networks (PIN, see [46] for seminal definitions); iii) MicroRNA gene (target) networks. In particular, PIN can employ stochastic network inference approaches (see [47–49], with regard to the resolution limit problem [50]), leading from modularity to sub-modularity through a variety of probabilistic methods comparatively evaluated and functionally validated at different scales (see also [51]).

A central point is that rather than the ability of identifying modules or clusters according to many possible algorithms, it is the accuracy with which modules and functional entities match that really matters, especially when multiple network layers (biological processes, protein complexes, and pathways) are considered [52]. We generally add further dimensions to gain information; for instance, investigation with reference to cross-correlation patterns of microRNA-target co-expressed profiles.

In the context of the *RNA-Seq* features, the identification (and diversity) of alternative-spliced isoforms [53] is very important. In particular, when considering the translation of AS-generated transcripts into proteins, it is natural to consider the *RNA-Seq* expression profiles as possible drivers of an in-depth exploration of PIN regions and detection of active sub-networks [54]. At both whole- and local sub-interactome scales such modules will be characterized by high connectivity densities and significant differential co-expression. This in turn will allow inference on condition/time-dependent regulators integrated within the network to assess the potential of driving modularity. From the design of differential regulation maps, the focus should go to pathways and modules aimed at detecting specific pattern propagation dynamics of the related signals. By looking at the configurations of active modules/motifs, it will be possible to observe changes in the structure of either separated or overlapping modules, pathway cross-links with respect to both bottlenecks and bridges (depending on the relevance of the information flow crossing the node, or also on the specific role played in each module), and whether such changes can be characterized in transient or permanent terms (for instance, the constitutive components form the core of complexes that change dynamically due to module addition/subtraction). Finally, concerning a combined use of drugs derived by identifying critical sub-networks rather than individual genes, the candidate set of target proteins will be localized in specific network regions and monitored relatively to distinctiveness and specificity of topological signature.

Such detailed information from the modules, also integrated with additional records (clinical, therapeutic, etc.), should deliver a few advantages: a) Expanding the potential number of candidate targets

from inferring both network-dependencies and module-specific influences on dysregulated cellular functions; b) Increasing the probability of selecting novel target candidates instead of highly targeted proteins (i.e. multiple compounds targeting the same protein) or off-targets (i.e. those not directly perturbing the proteins involved in cancer [55]); and c) Suggesting prediction of protein connections by assigning weights based on network robustness and stability properties measured at local rather than global scale.

12. Integrative view

Transcriptomics can provide a rich ground for causal inference, i.e. helping to elucidate the possible causes behind the changes of the organization assessed at network topology level. The correlation between transcriptional and protein network profiles is not completely predictable, as control/regulation mechanisms can be active at both levels. Several different protein variants may be encoded by genes, likewise post-translational modifications may occur in key proteins. However, scrutinizing the transcriptome profiles over time and coupling them with network configuration profiles to find correlated patterns, has a great potential, together with limitations. First, such coupling (or uncoupling) degree could be tissue-dependent. Then, regulatory roles of ncRNA or insight on possible gene extensions could be inferred from monitoring the changes in network configurations at both module and pathway scales.

Many transcriptionally active regions are currently not annotated [53], and novel discoveries are regularly coming out, as it was shown in [56] by the choice of the brain tissue as a source of evidence for pseudogenes (annotated or not). Notably, this study has revealed that under both normal (data source: Illumina Human Body Map 2.0 Project on transcription profiling computed on the basis of high-throughput sequencing of both individual and a mixture of sixteen normal human tissue RNA, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611) and disease (data source: TCGA — The Cancer Genome Atlas — http://cancergenome.nih.gov/cancersselected/lowergradeglioma, i.e. data from samples of lower grade glioma) conditions the obtained quantifications were in substantial amount.

By examining the distribution of intergenic reads (reads mapped between currently annotated gene, i.e., reads mapping outside the furthest 5′ and 3′ exons for every gene) across the genome, any densely mapped region could become a good target. However, reads mapping to intergenic locations tend to fall near annotated genes, suggesting that many annotations may require revision. This sort of prediction power underlying the transcriptome could be monitored through the dynamics transferred to networks and reflected into new associations (aggregation of separate interactions) and dissociations (break down) of modular structures. Intuitively, the natural quantity to monitor is the degree of participation in network activities, assuming that each module has a functional (i.e. biologically relevant) value. Thus, one goal is to check how differential conditions affect the participation to modules, and both measures of centrality and vertex–vertex distances offer insight on dynamics inducing a re-positioning of vertexes in the network.

13. Concluding remarks

Transcriptome complexities require, before proceeding to their in-depth characterization in any particular application, the following treatments: 1) The ability to perform deconvolution of transcript data structures by ad hoc computational statistics and machine learning approaches to make effective the targeting of structural entities, ncRNAs and gene isoforms (AS transcripts); 2) The examination of AS dynamics and development of isoform models with multiple features; and 3) A dynamic enhancement of the expression profiles of various types of transcript entities by considering time and space at both experimental measurement and computational analysis levels.

Challenging data structures call for innovative inference methods to deal with the complexities characterizing human transcriptomes under different conditions. The following objectives are set to achieve high-impact results:

i) Identify classes of problems with complexities for which innovative methods are needed;
ii) Design novel data-intensive inference approaches beyond agnostic learning and heuristic algorithms according to a well-defined model framework; and
iii) Provide proof-of-principle studies to establish legitimacy for the approaches.

Acknowledgments

The author thanks his colleagues at the University of Miami for fruitful discussions on the topics addressed in this review. The author is also grateful to three reviewers and to the Editor for their remarks and suggestions, which led to an improved paper.

References

[1] Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods 2011;8(4 s):S6–S11.
[2] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNASeq. Nat Methods 2008;5(7):621–8.
[3] Wang Z, Gerstein M, Snyder M. RNASeq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:5763.
[4] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G. Transcript assembly and quantification by RNASeq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28(5):511–5.
[5] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinforma 2010;26(4):493–500.
[6] Tarazona S, Garcia-Alcalde F, Dopazo Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Gen Res 2011;12:2213–23.
[7] Engl HW, Flamm C, Kugler P, Lu J, Muller S, et al. Inverse problems in systems biology. Inv Probl 2009;25:123014.
[8] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. NAR 2012;40(10):4288–97.
[9] Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. NAR 2014;42(11):e91.
[10] She Y, Hubbell E, Wang H. Resolving deconvolution ambiguity in gene alternative splicing. BMC Bioinforma 2009;10:237.
[11] Gonzales-Porta M, Calvo M, Sammeth M, Guigo R. Estimation of alternative splicing variability in human populations. Gen Res 2012;22:528–38.
[12] Jiang T, Wong WH. Statistical inference for isoforms expression in RNA-Seq. Bioinforma 2009;25(8):1026–32.
[13] Hiller D, Jiang H, Xu W, Wong WH. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinforma 2009;25:3056–9.
[14] Salzman J, Jiang H, Wong W. Statistical modeling of RNA-Seq data. Stat Sci 2011;26(1):62–83.
[15] Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. J Comput Biol 2011;18(3):305–21.
[16] Xia Z, Wen J, Chang CC, Zhou X. NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. BMC Bioinforma 2011;12:62.
[17] López-Ruiz R, Sañudo J, Romera E, Calbet X. Statistical complexity and Fisher–Shannon information: applications. Book ch. 4 , In: Sen KD, editor. Statistical complexity. 1st ed. Springer Books; 2011. p. 65–127.
[18] Chen R, Mias GI, Li-Pook-Than J, Jang L, Lam HYK. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148:1293–307.
[19] Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods 2014;11(6):637–40.
[20] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11(7):740–2.
[21] Donoho D. Compressed sensing. IEEE Trans Inf Theory 2006;52:1289–306.
[22] Candes E. Compressed sampling. Proc Int Congr Math. Madrid, ES: European Mathematical Society; 2006. p. 1433–52.
[23] Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 2009;12:833–44.
[24] Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ. The reality of pervasive transcription. PLoS Biol 2011;9(7):e1000625.
[25] Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer? Hum Mol Genet 2010;19(R2):152–61.
[26] Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long non-coding RNA pairs in the developing brain. PLoS Genet 2007;5:e1000617.
[27] Guttman M, Amit I, Garber M, French C, Lin MF. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.

[28] Khalil AM, Guttman M, Huarte M, Garber M, Raj A. Many human large intergenic noncoding RNAs associate with chromatin modifying complexes and affect gene expression. Proc Natl Acad Sci 2009;106:11667–72.
[29] Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A. Long noncoding RNAs with enhancer-like functions in human cells. Cell 2010;143:46–58.
[30] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915–27.
[31] Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 2010;42(12):1113–7.
[32] Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011;477(7364):295–300.
[33] Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol 2012;9(6).
[34] Johnson JM, Edwards S, Shoemaker D, Schaadt EE. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 2005;21:93–102.
[35] van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most 'dark matter' transcripts are associated with known genes. PLoS Biol 2010;8(5):e1000371.
[36] Mattick JS. The genetic signature of noncoding RNA. PLoS Genet 2009;5:e1000459.
[37] Mercer TR, Dinger ME, Mattick JS. Long noncoding RNAs: insights into functions. Nat Rev Genet 2009;10:155–9.
[38] Capobianco E. Entropy embedding and fluctuation analysis in genomic manifolds. Commun Nonlinear Sci Numer Simul 2009;14:2602–18.
[39] Capobianco E. Gene feature interference deconvolution. Math Biosci 2010;227:136–46.
[40] Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol 2012;8 [art. 565].
[41] Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004;36:1090–8.
[42] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3:140.
[43] Su J, Yoon BJ, Dougherty ER. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. PLoS One 2009;4(12):e8161.
[44] Pawson T, Linding R. Network medicine. FEBS Lett 2008;582:1266–70.
[45] Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 2009;27(2):199–204.
[46] Vidal M. Interactome modeling. FEBS Lett 2005;579:1834–8.
[47] Newman MEJ, Leicht EA. Mixture models and exploratory analysis in networks. Proc Natl Acad Sci 2007;104(23):9564–9.
[48] Marras E, Travaglione A, Capobianco E. Sub-modular resolution analysis by network mixture models. Stat Appl Genet Mol Biol 2010;9(1) [Art 19].
[49] Marras E, Travaglione A, Capobianco E. Protein interactomic manifold learning. J Comput Biol 2011;18(1):81–96.
[50] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc Natl Acad Sci 2007;104(1):36–41.
[51] Capobianco E, Marras E, Travaglione A. Multiscale characterization of signaling network dynamics through features. Stat Appl Genet Mol Biol 2011;10(1) [Art 53].
[52] Marras E, Travaglione A, Chaurasia Futschik M, Capobianco E. Inferring modularity from human protein interactome classes. BMC Syst Biol 2010;4:102.
[53] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. Gen Res 2008;18:1509–17.
[54] Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 2013;14:719–32.
[55] Yildirim MA, Goh K, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol 2007;25(10):1119–26.
[56] Valdes C, Capobianco E. Methods to detect transcribed pseudogenes: RNA-Seq discovery allows learning through features. In: Poliseno L, editor. Pseudogenes functions and protocols. Methods in molecular biologyHumana Press; 2014. p. 157–83.