ABSTRACT

Biomarkers are cornerstones of healthcare. Their applications include disease diagnosis and prediction, indication of disease progression, and monitoring of treatment response. Despite significant efforts that have identified thousands of potential biomarkers, their translation into clinical practice remains poor: less than two approvals per year across all diseases. In part, this inefficiency arises from both experimental and statistical limitations on the biomarker discovery pipeline. Widely used mass-spectrometry approaches suffer from sample throughput whereas targeted approaches such as single and multiplex immunoassays suffer from the number of interrogated proteins and antibody quality. On top of the technological limitations, the current single-biomarker-to-single-disease approach does not capture the multifactorial nature of complex diseases. A disease can be originated from diverse pathological mechanisms, while a pathological mechanism can lead to diverse diseases. Hence, mechanism based biomarker discovery aims to identify signatures that capture the diversity of the disease’s origin and deliver more precise diagnostic and predictive information.

INTRODUCTION

Defined as objectively measurable indicators of specific biological conditions, particularly the ones related to disease [1], biomarkers are constantly expanding their influence in modern medicine. Some recently qualified biomarkers include urinary clusterin as an indicator of drug-induced kidney injury [2], troponin T as a marker of cardiac damage [3], or galactomannan for the diagnosis of invasive aspergillosis [4]. Biomarkers are essential tools in health-care and are widely used for the diagnosis, prediction, and prognosis of disease. Biomarkers are also becoming essential tools in reducing the failure rates in clinical development by assessing drug efficacy, toxicity, and guide patient recruitment for targeted therapies [5].

There are many types of biomarkers. Imaging-based biomarkers includes Magnetic Resonance Images (MRI), Positron Emission Tomographies (PET), X-ray, and Computed Tomographies (CT). On the other hand, molecular biomarkers include genes [6], transcription factors [7], proteins, metabolites [8], lipids, and glycans. Biomarkers can also be naïve ones such as weight, body temperature, the heart rate for fibromyalgia syndrome [9], or the prothrombin time for anticoagulant activity [10]. The present review focuses on molecular biomarkers since this is the main source of information for obtaining the mechanism of a disease. Molecular biomarkers include DNA, RNA, metabolites, and protein biomarkers measured from different biological fluids (plasma, serum, urine, synovial fluid, etc.). For monitoring human health, measuring protein biomarkers in blood is considered a very attractive solution because of the simplicity of obtaining the sample, and because the pathology of almost every body tissue can affect the blood proteome [11].

STATE OF THE ART

The traditional mass-spectrometry (MS) to ELISA protein biomarker pipeline is a data-driven
approach that combines the untargeted, broad, and statistically weak MS search with the targeted, narrow, and statistically strong ELISA validation. It consists of three main steps [1,12]: (i) The discovery step identifies hundreds to thousands of candidate biomarkers by quantifying a few samples (10s) coming from proximal fluids and using untargeted MS analysis. (ii) The qualification and verification step confirms the differential expression of selected candidate biomarkers in human plasma and starts assessing some of their indicators (specificity, sensitivity, etc.) by quantifying from tens to hundreds of analytes via targeted MS. The number of patients introduced in this step ranges the 10s-100s. (iii) In the validation and approval step, immunological affinity assays are developed for the 1-10 most sensitive and specific biomarkers, which are then validated in a large number of samples (usually 1000s), and approved for clinical use. Using this pipeline successful biomarkers have been cleared and approved. Examples are Galectin-3 for myocardial fibrosis [13] or serum tryptase for clinical hematology [14]. Despite this biomarker pipeline led to the discovery of 1000s of candidate biomarkers in the last 20 years, their translation into clinical diagnostic assays remain very poor: less than 2 approvals per year across all the diseases [15]. This difficulty and controversy draws special attention in expression markers in cancer where many biomarkers have been discovered, but very few have been proven successful in the clinic. The MammaPrint assay is a 70-gene signature assay used to personalize the treatment for patients with breast cancer [16] which probably represents the most successful expression-based biomarker(s) that is currently in use in clinical trials. Several factors have contributed of this poor performance [17]:

- Biomarker discovery needs statistical power and thus large number of samples. Sample throughput limitations of the MS technology at the early phases [1] combined with the traditional one-biomarker-one-disease approach cannot address co-morbidity factors (i.e. biomarkers present in co-occurring conditions) resulting in low specificity. Low sample number also fails to provide the statistical power required for high biomarker sensitivity (ability to identify positive samples across all samples).
- Diseases are often multifactorial, caused by multiple underlying pathological mechanisms. For example, in cancer research, cancer can originate from diverse cellular mechanisms known as hallmarks [18], while a cell survival mechanism can lead to diverse cancer types. In the same way, a pathological mechanism can lead to diverse diseases.
- The transition from candidate biomarkers identified via mass-spec (early phases of the pipeline) into affinity assays (late phases of the pipeline) is time consuming and often inefficient due to the lack of high-quality antibodies [19,1].
- Biomarker discovery poorly exploits the available knowledge on the underlying mechanism of disease pathology and drug action, thereby failing to accelerate the process. This information could provide better guidance on which candidate biomarkers (provided by the discovery phase) to proceed in the following tedious and expensive phases of biomarker discovery.
- The challenging nature of the serum proteome and the variability of protein abundance (from albumin with reference range 35-55 g/L to the billion-times less concentrated proteins) make accurate protein quantification hard. [12].

Although many of the above challenges might be solved with new proteomic technologies, the
multifactorial nature of the diseases requires a fundamental shift that currently takes place in the biomarker discovery which takes into consideration the diversity of the pathological mechanisms. On this front, the advancements in understanding disease mechanisms based on gene expression, sequencing and proteomic data are in the process to be incorporated on the biomarker discovery pipeline.

Pathway-based biomarker discovery does not build solely on finding differentially expressed proteins, but rather uses holistic approaches that integrate experimentally obtained data with biological knowledge (pathways) to interpret the results and gain biological insights about the mechanism behind those results [20]. Though large changes in crucial individual proteins can have important consequences on the phenotype, small changes in functionally related groups of proteins (clusters) can also have significant effects [21]. Thus, pathway-based biomarker discovery comes early in the discovery phase and incorporates the omics-derived biological knowledge in order to deliver statistically significant biomarker candidates that can be used to identify both the disease and its mechanisms. In an effort to incorporate the different methods presented in the literature, four elements are required for pathway-based biomarker discovery (Figure 1): (i) multi-omics data gathering, (ii) knowledge base (pathways, clinical data, and drug information), (iii) computational analysis, and (iv) experimental validation.

Figure 1 – Schematic overview of the traditional and current biomarker discovery pipelines. The traditional pipeline uses untargeted experimental methods (e.g., untargeted MS) for candidate proposal; whereas the current pipeline exploits the plethora of multi-omics data available in public repositories and the biological knowledge about the functionalities of different genes, proteins, etc. (knowledge-base) for the same purpose. The verification process of the traditional pipeline elects the most promising candidates suggested in the previous step using targeted approaches (e.g., targeted MS). In both cases, final candidate biomarkers are validated using immunoaffinity assays (e.g., ELISA) before going to clinical trials.

DATA GATHERING

The advancement of “-omics” assays has boosted the use of high-throughput technologies
(MS, NGS, etc.) which generate large quantities of data leading to a plethora of information about genes, transcripts, metabolites and proteins that are available in many public domain repositories [22]. Current pipelines include in-house data as well as publicly available data opening the door to in-silico approaches for the first phase of biomarker discovery. Data mining methods can exploit already available (and publicly accessible) data from similar studies and use in-house analytic methods to obtain the biological insights and the mechanism behind the condition under study. Data mining can be done either by using some of the many available tools and platforms [23-26], as well as searching for similar studies in any of the many public repositories accessible online like the Gene Expression Omnibus (GEO) source (https://www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) for genomic data, or The PRoteomics IDEntifications (PRIDE) source (https://www.ebi.ac.uk/pride/archive/) for proteomic data, or the Human Metabolome Database (HMDB) source (http://www.hmdb.ca) for metabolomic data. Currently, there exist more than 2,000 multi-omic data repositories containing molecular data which can be accessible via the internet [20]. Consequently, both, the type of data (proteins, genes, etc.) and the technology (technologies) to be used, play a major role in current biomarker discovery pipelines.

Proteomic assays can be classified into two main technologies: Mass Spectroscopy (MS)-based proteome profiling, and affinity multiplexing assays. Recent advances in both technologies need special attention as they provide opportunities to enhance the discovery and clinical translation of serum protein biomarkers. Early discovery phases require wide proteome coverage (large number of analytes per sample, a parameter denoted as “multiplexability”) and thus, MS approaches are preferred despite their very low throughput (number of samples quantified per day) [27]. ELISA-based immunoassays are predominantly utilized in biomarker validation due to their high throughput despite their limited multiplexability. Clearly, technological advancements are required to bridge the gap from discovery-to-validation.

Traditional MS-based approaches can quantify thousands of proteins per sample, however, the number of samples that can quantify per day is limited which does not allow more than a few dozen samples to be measured in the discovery phase [28]. Moreover, their repeatability in the detection of low-abundance proteins is far from perfect due to the use of untargeted experimental methods [1]. The low throughput of MS approaches also prevents their utilization in the later phases of the pipeline, where hundreds or thousands of samples need to be quantified. MS combined with fractionation (e.g. LC-MS/MS) enhances its resolving power at the lower concentration range, however at significant cost of throughput [29]. To overcome the limitations of traditional (untargeted) MS approaches, new targeted MS approaches, such as MRM-MS, have recently gained popularity, particularly in the validation phase [30]. These methods utilize isotope-labeled standard peptides homologous to peptides of the “target” protein in the sample [31]. MRM-MS offers specific, accurate, and wide dynamic range quantification for an extensive range of targeted protein concentrations in a significantly higher samples throughput. However, their application is complex and requires substantial optimization. The newer SWATH strategy provides higher throughput by utilizing a wider MS/MS window [32], however its application in real clinical samples is challenging.

As for transcriptomic assays, next-generation sequencing (NGS) technologies, specifically RNA-
Seq, have promoted enormous advances in our understanding of the transcriptome over the last years providing more accurate measurements of transcriptomic levels than other technologies [33]. Several recent studies use RNA-Seq data for biomarker discovery in different types of cancer [34,35]. However, hybridization or sequence-based approaches like RNA microarrays are still proving successful for transcriptomic-based biomarker discovery. For example, the well-known Affymetrix Human Genome U133 Plus 2.0 Array was recently used for lung-cancer diagnostic in saliva samples [36].

Regardless the technology of choice, functional correlation of the measured biological entities and modelling of biological pathways is required for mechanistic understanding of disease/drug function and predicting cell behavior. Several types of models have been proposed over the years to capture the signaling behavior of various cells [37]. The choice of appropriate models depends largely on the type and size of available data and the required level of mechanistic detail. In this context, pathway-based biomarker discovery introduces the concept of knowledge-base in order to bring functional, correlational, contextual, and causal information about the biological entities participating in these models.

KNOWLEDGE BASE

Traditional biomarker discovery focuses on finding individual biomolecules that are differentially expressed in two distinct biological conditions. This data-driven approach does not take advance of the wealth of information that can be hidden behind each biomolecule and can find in knowledge bases connections to other biomolecules, diseases, pathways, clinical trials, or drugs. On this front, mechanism-based discovery centers on the biological mechanism behind groups of functionally correlated entities that are altered between the two conditions. Therefore, knowledge bases contain the required information to transform a data-driven approach for biomarker discovery to pathway-based biomarker discovery [38] that contains biological interactions, disease mechanisms, and functional annotations of the entities under study (Figure 2).

A common feature of every knowledge base is that it requires annotation of the biological function of the genes, metabolites, proteins, etc. Over the last years, a big effort has been made collecting this type of biological knowledge. The Kyoto Encyclopedia of Genes and Genomes (KEGG) resource (http://www.genome.jp/kegg/) is a reference database for correlating biological entities to their function. Reactome (http://www.reactome.org) and OmniPath (http://omnipathdb.org/), are another two literature-curated, top-level, databases for human signaling pathways and reactions. PANTHER (http://pantherdb.org), is a very well-known curated database containing information about gene and protein function. In this way, Systems Biology has emerged from the need to understand, integrate, and interpret the resulting big datasets. With the traditional biomarker discovery pipeline, the limitation resided on collecting the data; with current pipelines jumps to integrate and analyze such data. It is to note that the different nomenclatures for the different biological entities for the different databases make such integration challenging. The Open Biomedical Ontologies (OBO) consortium is an initiative trying to solve this problem [39]. In the same line, the Molecular Signatures Database (MSigDB) source (http://software.broadinstitute.org/gsea/msigdb/), is pursuing the same objective.
COMPUTATIONAL ANALYSIS

The current single-biomarker-to-single-disease approach does not capture the multifactorial nature of complex diseases. Molecular entities rarely act alone; therefore, single genes/metabolites/proteins fail on providing biological insight about the underlying mechanism of the condition under study. A disease can be originated from diverse pathological mechanisms whereas a pathological mechanism can lead to diverse diseases. This, united to the complexity of high throughput data analysis, has favored pathway analysis to become the method of choice during the last 15 years. Pathway analysis integrates single genes/proteins from large lists into smaller sets, including in each set entities that are known to interact with each other. This simplification not only increases the explanatory power of individual entities but also reduces the complexity of the later analysis [40].

Khatri et al. [38] brightly reviewed the different methods developed for pathway analysis during the last years. Briefly, three generations of methods have been developed, each of them addressing the limitations of the previous generations. First generation methods use Over-Representation Analysis (ORA) where the implication of a pathway is calculated via the fraction of genes contained in the pathway among the set of differentially expressed genes. Second generation methods, Functional Class Scoring (FCS) methods, support the idea that small changes in a set of genes participating in the same pathway can have a biological impact as strong as large changes in individual genes. Finally, third generation methods, Pathway Topology Based (PT-based), are FCS methods that also exploit information about what gene products interact with each other as well as how and where they do interact.

As an example, Gene Set Analysis (GSA) methods aim to dodge the evaluation of the activity of individual genes by integrating such genes into groups and evaluate the activity of the consequent clusters. In recent years, GSA has gained attention because it gets rid of the threshold based methods used in Individual Gene Analysis (IGA) [19]. Gene Set Enrichment Analysis (GSEA) [41], identifies statistically significant differences in pre-annotated gene clusters between two biological states. A large number of methods have been proposed for GSA, however, there is not yet a unique reference method for this stage of the analysis. Therefore, some methods have been developed integrating different GSA approaches. For example, the R Package Piano [42] collects a series of GSA methods and uses a consensus scoring approach for the final result. It is to note, that for GSA be effective, it needs previous annotation of all the gene clusters containing what genes are included on each cluster and what biological function do they carry out as a group. Such information can be found in several databases as the previously mentioned MSigDB [43]. However, there are a lot of sequence fragments whose function is still unknown making IGA a better approach when this information is missing.

Gene Scoring Based Methods (GSBM) first identify few genes important for the condition under study and then assess a score to each other gene based on proximity or protein-protein-interaction (PPI) networks [42]. The objective is to prioritize genes (clusters) more closely related to the disease and not consider all the genes equally important. This same concept has been used to screen drug efficacy [44]. All these methods use prior knowledge of the disease to select initial candidates or reference genes and then apply topological information to
prioritize. The Online Mendelian Inheritance in Man (OMIM) resource (http://www.omim.org), is a public repository containing relations between genes and diseases.

Theoretically, all the methods described above are independent of the type of data used (genes, proteins, metabolites, etc.), as far as these are linked to pathways via the knowledge base. Consequently, a well-defined knowledge base is crucial to ensure the success of the biomarker discovery approach.

Figure 2 – Schematic overview of the current biomarker discovery pipeline and available online repositories and methods. The current biomarker discovery uses both in-house and in-cloud data from different types of biological entities, i.e., genes, proteins, metabolites, etc. as well as functional, topological, and causal knowledge about such entities in order to reduce the complexity and increase the explanatory power of the latter analyses. Note that only some of the most important repositories/methods have been included.

VERIFICATION AND VALIDATION

As previously defined, verification and validation steps are oriented to the candidate biomarkers identified during the discovery phase, and therefore, utilize targeted experimental methods which are potentially faster. Affinity multiplexing assays have recently emerged as the reference proteomic method for verification and validation phases of the biomarker discovery pipeline. They rely on an affinity molecule (antibody, aptamer, affimer, somamer) that recognizes specific epitopes on the protein of interest with high affinity and specificity. Their main advantage is their ability to quantify large numbers of samples, required for the validation of the sensitivity and specificity of a candidate biomarker. Affinity multiplex assays are sensitive (especially the ELISA sandwich format), require small volumes of sample, and minimal sample preparation compared to MS. Their key limitation is their reliance on antibodies, which many times suffer for cross-reactivity issues and questionable performance [45,46]. Affinity multiplex assays can be categorized based on their format (i.e. planar vs. in-
suspension/bead-based) and their ELISA format: sandwich ELISA, single antibody/direct labelling, sample-down/reverse phase (for a review see [47]).

The analytical methods used at this stage are focused on assessing specificity (ability to identify negative samples across all non-diseased samples), sensitivity (ability to identify positive samples across all diseased samples), bias (representation of all the cases, related to sample size), and robustness (consistency of the results across test conditions) of the candidates derived from the previous phase. Whereas bias and robustness mostly depend on the sample size selected for the verification and validation steps, specificity and sensitivity rely on the analytic method chosen. Since the final objective is to select a panel of biomolecules that classify patients between two different biological conditions, verification and validation approaches use machine-learning methods (both supervised and unsupervised) to select the optimum set of biomarkers for discrimination. Even if Receiver Operand Characteristic (ROC) curves remain the gold standard for this type of analysis, Artificial Neural Networks have been proven very useful for cancer prediction problems [48,49], and Support Vector Machine (SVM)-based methods, are commonly used for classification and regression analysis. New developments in machine learning such as deep learning have improved the modelling of highly abstract problems by introducing several processing layers constituted of numerous linear and non-linear transformations. These methods have improved, among many other categories such as speech and visual recognition, the state-of-the-art in drug discovery and genomics [50]. Other families of methods used at this stage can be: genetic algorithms, decision trees, or Bayesian classifiers. As a successful example, Random Forest, a machine learning algorithm that uses classification and regression tools for classification [51], has been recently proven useful for metabolomics-based biomarker discovery in rats’ urine [52].

CONCLUSION

Although some successful biomarkers have progressed until clinical trials during the last years, the traditional biomarker pipeline does not capture the underlying biological mechanism behind complex diseases, what has been translated to poor validation performance. This factor, along with the growth of multi-omic data technologies and the improvement of functional knowledge of genes and proteins, has originated pathway-based biomarker discovery to gain significant importance during the last decade.

Pathway-based biomarker discovery does not build on finding individual biological entities which are differentially expressed in two distinct biological conditions, but rather uses holistic approaches that integrate experimentally obtained data with biological knowledge (pathways) aiming to discover the underlying mechanism behind groups of functionally correlated entities and gain biological insights that can help the interpretation of such results. It seems clear that, due to their nature, multifactorial diseases will benefit from pathway-based biomarker discovery much more than others. In this context, DNA biomarkers in blood for different types of cancer (known as liquid biopsy for cancer) can largely benefit from pathway-based biomarker discovery due to the possibility of tracking the mutations in circulating tumor cells (CTCs) and predicting the cancerous nature of circulating free DNA (cfDNA) in the blood, aiding the estimation of the risk of metastatic progression, the stratification of therapies, heterogeneity of tumors, etc. [53, 54].
Pathway-based biomarker discovery benefits from well-defined functional annotation of biological entities. However, not the function of all genes/proteins is equally well understood restricting the analytic methods. Despite these limitations, improvements in experimental technologies, data annotation, and computational tools provide new opportunities on biomarker discovery for reducing the failure rates in clinical development as well as for the diagnosis, prediction, and prognosis of multifactorial diseases.

ACKNOWLEDGMENTS

This work was supported by the European Union Horizon 2020 grant SyMBioSys MSCA-ITN-2015-ETN #675585 that provided the financial support for A.A.

REFERENCES


