1 Introduction and aim of the study

Bacillus megaterium is a rod-shaped Gram-positive soil bacterium, which was first discovered in 1884 by Anton de Bary [1, 2]. It got its name from the Greek “megatherium” for “big beast” because of its enormous size of up to 2.5 x 2.5 x 10 µm^3 (**Fig. 1.1**). Within the bacterial kingdom, these remarkable dimensions have propelled it to model organism of choice for single-cell analysis and investigation of cell structures and protein localisation [3, 4]. Cell wall synthesis, sporulation, bacteriophages and biochemistry of Gram-positive bacteria have been, for instance, widely studied using *B. megaterium* [5-8]. Besides its main natural habitat, the soil, its proficiency to metabolise a large range of carbon sources and its high osmotic tolerance has enabled *B. megaterium* to colonise varied ecological niches such as sea, industrial wastewaters and food products like honey or dry meat. This versatility and its ability to produce a large range of industrially relevant products have progressively made it an essential bacterial cell factory.

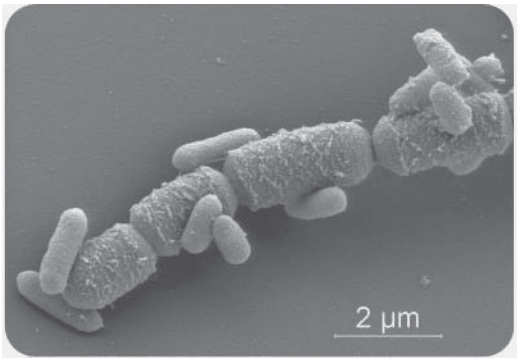

Figure 1.1: Scanning electronic microscope (SEM) pictures of *B. megaterium* (2.5 x 2.5 x 10 µm^3) and *E. coli* (0.5 x 0.5 x 2 µm^3) (M. Rohde; Helmholtz-Zentrum für Infektionsforschung GmbH, Braunschweig, 2006).

A decisive step towards the widespread use of *B. megaterium* in the industry is undoubtedly the introduction and development of a xylose inducible promoter system for heterologous plasmid-based protein production by Rygus and Hillen (**Fig. 1.2**) [9]. The natural system consists of the genes *xylA*, *xylB* and *xylT* encoding xylose isomerase, xylulokinase and xylose permease, respectively [10]. Divergently to these genes, the gene *xylR* encodes the repressor XylR regulated by P_{xylA}. In the absence of xylose, the repressor binds the operator regions O_L and O_R of the promoter P_{xylA} and transcription of all genes downstream cannot be initiated. On the contrary, upon addition of xylose, the repressor protein XylR binds the xylose, undergoes a conformational modification and can no longer bind the operator regions. As a consequence, RNA-polymerase mediated transcription of the *xyl*-operon is derepressed and increased by 150 times in comparison to the inhibited state. Apart from this main control system, two additional mechanisms regulate the expression of the xylose operon when glucose is present. On the one hand, glucose enhances the binding affinity of the catabolite control protein A (CcpA) for the catabolite repression DNA-element *cre* located in the gene *xylA* and thereby hinders the proper transcription of the whole operon. On

the other hand, assimilation of extracellular glucose generates significant intracellular amounts of its phosphorylated counterpart glucose-6-monophosphate, which can outcompete xylose in binding repressor protein XylR and thus prevent operon transcription. Overall, these two mechanisms account for a 14 times lower transcription level in the presence of glucose. The catabolite responsive element was therefore subsequently removed on the corresponding vector system to obtained a system suited for recombinant protein production using glucose as carbon source and xylose as inducer [11]. Later, further optimisation of the promoter, the ribosome-binding site and untranslated 5' mRNA region (5'UTR) resulted in an up to 12-fold improvement of the system global efficiency [12].

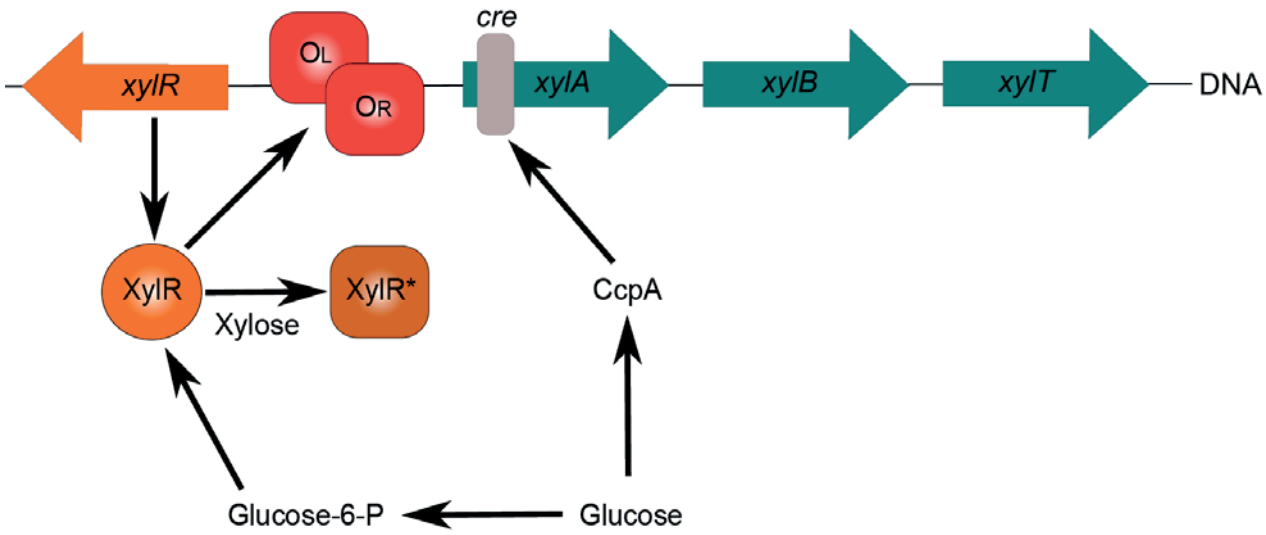

Figure 1.2: Regulation of the xylose-operon in *B. megaterium* – CcpA: catabolite control protein A, ***cre***: catabolite response element, **O_L/O_R**: operator region of the xyl-promoter, ***xylA***: xylose isomerase gene, ***xylB***: xylulokinase gene, ***xylR***: xylose repressor gene, **XylR**: active xylose repressor protein, **XylR***: inactive xylose repressor protein

In addition to its stable plasmid replication system, *B. megaterium* presents several other advantages in comparison with traditional industrial workhorses such as *Escherichia coli* or *Bacillus subtilis*. Firstly, it exhibits a high secretion capacity combined with the lack of an outer membrane [13]. So secreted products can directly be collected from the supernatant. Secondly, whereas several alkaline proteases are produced by *B. subtilis*, none of them were found in *B. megaterium* and produced exoenzymes accordingly show a remarkable stability [14]. Thirdly, the lack of endotoxins in *B. megaterium* and its non-pathogen status makes it an ideal production host for pharmaceutical and food applications, for which safety issues often impose expensive downstream processing otherwise.

At first, only unaltered wild-type strains were used for the production of a limited number of compounds comprising vitamin B_{12}, α- and β-amylases, xylanase, penicillin G acylase and polyhydroxybutyrate (PHB) but the introduction of the plasmid-based expression system has widen the product spectrum to varied recombinant proteins and sophisticated compounds such as

antibody fragments, glycosyltransferase (levansucrase, dextransucrase) and the green fluorescent protein (GFP). The latter being a particularly useful model protein for assessing the promoter efficiency or for monitoring the impact of process parameters on recombinant production [15, 16].

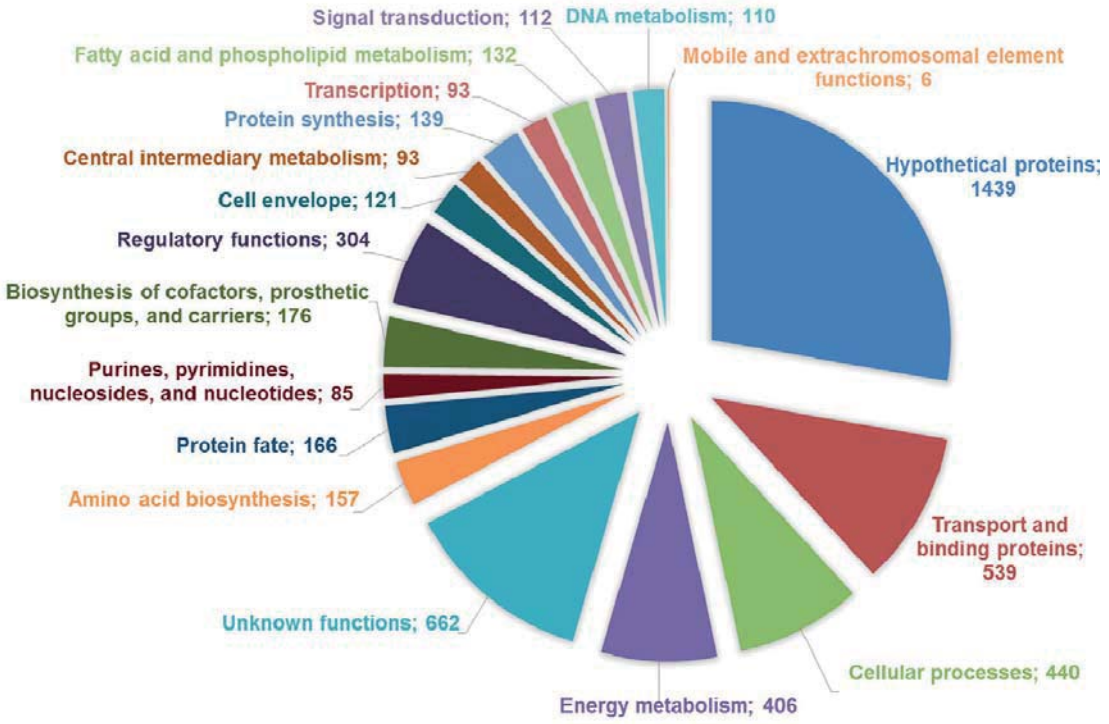

Figure 1.3: Classification of genes from *B. megaterium* DSM319 into TIGR role categories – Functions were attributed according to the sequencing of its complete genome by Eppinger et al. [17].

Recently, the sequencing of the complete genome of three different strains and the fast development of dedicated omics-techniques have furthermore laid the foundations for an in-depth understanding of its metabolic behaviour and opened up new possibilities towards its rational genetic modification (**Fig. 1.3**) [17-20]. This system-wide approach should in term enable the elucidation of all metabolic and regulatory steps involved in the production of a given substance and predict subtle targets for metabolic engineering.

This study takes place in this context of continual improvement of *B. megaterium* as a production host and was set out to get a better comprehension of its metabolic behaviour and of regulatory mechanisms involved in response to two industrially relevant issues, namely temperature and osmotic stress. Taking advantage of the recent technical developments of systems biology, system-wide response to these two adverse conditions shall be assessed for the first time in this organism in a multi-omics study including transcriptome, proteome, metabolome and fluxome analyses. For the latter, condition-specific macromolecular biomass compositions shall be determined and corresponding precursor demands integrated in a brand new model. Results obtained from the different omics-techniques shall then be analysed separately, combined together and with gathered physiological data to provide a functional understanding of metabolic adaptation of cells responding to temperature (between 15 and 45°C) and osmotic stress (mimicked with up to

1.8 M NaCl). Finally, potential genetic targets shall be identified using generated data sets and implemented to further optimise robustness and production characteristics of *B. megaterium*.

2 Theoretical background

2.1 Systems biology and omics technologies

2.1.1 Systems biology and its recent development

Life is a complex, multifaceted and evolutive process involving sophisticated and fascinating mechanisms such as tissue regeneration, immune response or thermal homeostasis. It is unfortunately an imperfect one as well, in which dysfunctions such as cell degeneration, hormonal disorders or memory loss may occur. In recent years, it has become obvious that no matter how meaningful the breakthroughs within the single fields of biology are, they will never be able to address this complexity and provide viable healthcare solutions if considered separately. Of course, it is in the first place of crucial importance to know of which biological components (genes, proteins, transcripts, metabolites, pathways) a living organism disposes and what the possible interactions between them are. However, since life is not static, it is even more important to unravel global regulation networks orchestrating those interactions *in vivo* and defining how biological components actually function together as a whole. From these considerations, systems biology emerged as a science willing to remodel the classical and segmented approach of biology into a highly interdisciplinary and informational one, where interaction and control dynamics between single biological layers would also be assessed (**Fig. 2.1**).

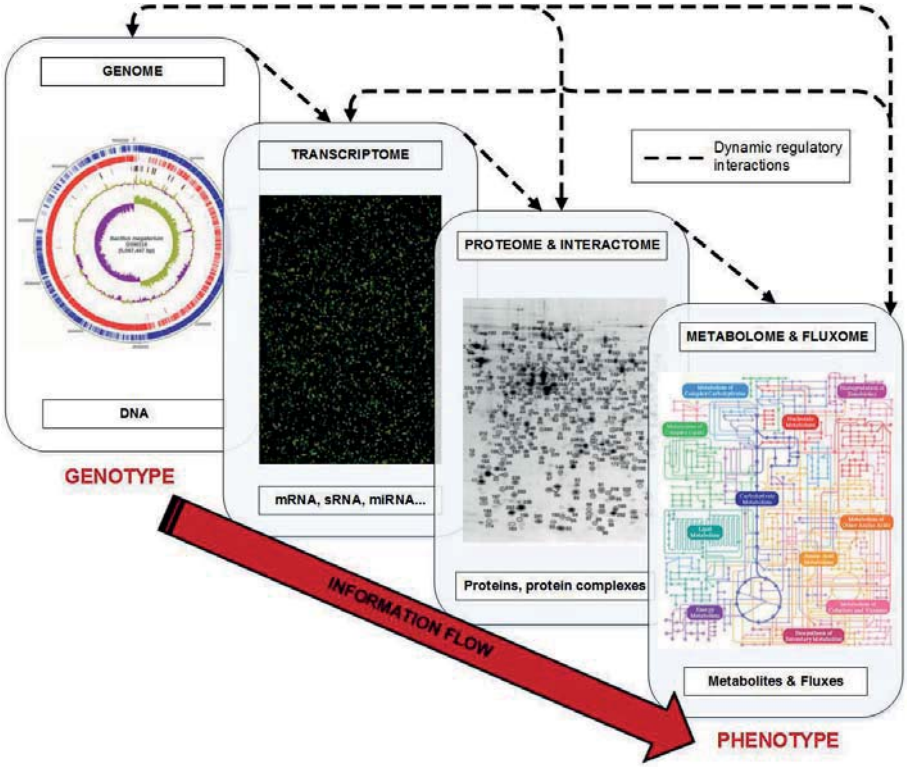

Figure 2.1: Architecture of cellular systems and interactions among the different functional layers – Dashed lines represent dynamic regulatory interactions between molecular species. Figure was adapted from [21] and [22].

Such a functional and system-wide comprehension was only made conceivable by the parallel fast development of high-throughput omics-technologies and advanced computational methods, which enabled the acquisition and processing of large amounts of experimental data. Indeed, to comprehend the global regulation of biological structures, systems biology systematically perturb organisms in various ways and records their reaction at different organisational levels, including gene expression and protein production, modification of metabolite pools and pathway utilisation. Collected data are afterwards integrated in global *in silico* models containing all known and hypothesised regulatory systems and contribute to their iterative refining by corroborating or rejecting initial model assumptions. As the generated data and underlying biological interplays are far too substantial and complex for human brains to deal with, computers arise progressively as the cornerstone of this new approach. They provide scientists with numerous databases indexing uncovered metabolic pathways, genetic information and interaction patterns but also with simulation tools able to confirm, discard or even suggest hypothesis that would otherwise not necessary be apparent to human beings [23]. Moreover, they are intensively employed in effective experimental design, thereby avoiding irrelevant analysis and reducing laboratory efforts needed to address specific issues. Another critical turning point for the boom of systems biology was the rapid development of automated and standardized genetic tools achieved within the framework of the human genome project (HGP), paving the way to fast sequencing, systematic gene deletion, insertion and mutagenesis [24]. After that, systems' perturbation could be performed not only by changing abiotic conditions but also through targeted modification of organisms' intrinsic capabilities and scientists could easily manipulate organisms to resolve specific regulatory pathways.

Although systems biology has already extended our knowledge of cell function and physiology in many ways, several barriers still prevent it from reaching its full potential. Firstly, measurement accuracy and coverage of actual devices remain insufficient to supply enough information for the complete determination of metabolic and regulatory properties of cellular systems. Development of even more efficient computational methods could partly compensate this problem but further technical advancements are inevitable. Secondly, the access to high-throughput and computational technologies is still limited due to their price and/or the level of expertise their operation requires. As institutes are usually specialized in only one or two domains, they cannot perform a system-wide analysis alone. Hence, the creation of solid research networks regrouping teams with complementary skills seems to be a prerequisite to widen the actual scope of systems biology. Lastly, no efficient pooling of collected omics-data has been implemented so far and information exchange between researchers at a global level is not trivial. Overcoming these challenges depends to a large extent on breakthroughs in other area such as computer science, biochemical engineering, physics or chemistry. Thus, systems biology arises as a strong driving force for scientific innovation.

In spite of still being in its infancy, systems biology has proven to be a promising area of research with a broad range of applications in both academic and industrial fields. Far beyond the single

understanding of life and its evolution, unravelling regulatory mechanisms gives us the keys to predict how genetic manipulations and induced metabolic interferences will affect phenotypes. Consequently, in the future, systems biology will undoubtedly play a central role in developing more effective therapeutic treatments with minimal side effects or also in improving bacterial cell factories. These new industrial workhorses will no longer be generated by random mutagenesis but rather rationally designed to be less stress-sensitive and less inclined to unnecessary by-products secretion, revolutionizing our common conception of bioprocess design in which production process must be adapted to bacteria and not the opposite [25, 26].

2.1.2 Genomics and Transcriptomics

Thanks to the fast progress of sequencing techniques achieved over the last three decades, genomes can now easily, swiftly and cheaply be sequenced. With more than a new bacterial genome completely sequenced every month, biological research has moved to a post-genomic era, where the gathered genetic information has to be organised into functional structures to depict the global dynamics of living cells [27, 28]. In this context, the identification and quantification of the complete set of transcripts present in a cell under given physiological conditions, referred to as transcriptomics, has proved to be a powerful approach to gain new insights into gene functionality and their regulation [29]. Historically, gene expression has first been locally analysed using Northern blot, where RNA transcripts from samples are first separated by electrophoresis and subsequently hybridised with labelled complementary probes [30]. Later, the discovery of reverse transcriptase, which converts mRNA into its complementary DNA (cDNA), has enabled the development of real-time reverse transcription polymerase chain reaction (qRT-PCR), the most sensitive technique presently available for quantifying RNA [31, 32]. However, qRT-PCR is a gene-specific procedure and monitoring gene expression levels at the genome scale with this technique would require a great deal of time and effort.

On the contrary, DNA microarray, a technology developed approximately thirty years ago, offers a straightforward and reliable way to identify and quantify the expression levels of a hundred thousand of genes simultaneously [33, 34]. To this end, DNA probes specific to parts of every single gene sequence of the investigated organism are either mechanically deposited or in-situ synthesised in the grid cells of a glass, plastic or nylon chip [35]. In parallel, RNA transcripts from given samples are purified, directly labelled or reverse-transcribed to their more stable cDNAs and labelled afterwards with fluorescent dyes. Subsequently, these labelled cDNA transcripts are hybridised to their DNA counterparts immobilised on the surface of the chip. After removing unbound transcripts by washing the array slide, labelled strands are excited using dye-specific wavelengths. The light emitted from each grid spot is captured in a scanner by a photo-multiplier tube (PMT) and converted into a digital image [36]. After algorithmic post-processing of this image including grid alignment for gene identification, spot characterisation (size, intensity, quality and outlier removal), background correction and intensity normalisation, the expression of a given gene is obtained from its corresponding spot intensities. Most of the time, microarray analyses are used

for direct comparison of gene expression between two samples (experimental vs. reference) and carried out as double-channel experiments, meaning that transcripts originated from samples are labelled with distinctive dyes (e.g. cyanines cy3 and cy5), hybridised on the same chip and their relative expression levels obtained by scanning the array at two wavelengths (e.g. red and green for cy3 and cy5, respectively) (**Fig. 2.2**) [37]. Since it is cheaper and does not need to be corrected for batch effects, this approach is often preferred to single-channel experiments, for which samples to compare are hybridised separately on two arrays using a single dye. However, if numerous samples need to be compared and thus the use of different microarrays is inevitable, single-channel experiments can be preferred to prevent aberrant samples from contaminating data derived from others and to get rid of eventual dye-related artefacts. It is therefore essential to choose the most appropriate experimental design with respect to the addressed biological issue to maximise the output of the analysis [34]. In this respect, it is also of outmost importance to define the number and nature of replicates needed to reach statistical relevance. While technical replicates tend to become superfluous as technology progresses, at least 3 to 5 biological replicates should be used for cDNA microarrays [38-40].

After completion of the microarray experiments, a tremendous quantity of information is available and the main challenge for researchers is to make sense out of these data. First, measured expression levels are normalised using either internal standards or statistical parameters such as standard deviation, mean and median values inter and intra arrays to improve comparability of microarrays [41]. To facilitate pattern discovery, data complexity is then drastically reduced by applying statistical filters that only retain genes whose regulation is significantly modified under the evaluated conditions. Typically, a cut-off value for gene expression is arbitrarily set and the pertinence of the resulting candidate selection is statistically assessed using various tests such as Student or Welch's t-tests, analysis of variance (ANOVA) and the false discovery rate (FDR) [42-46]. Finally, once the significance of the data is established, different clustering algorithms (hierarchical, k-means, SOM) can be applied to regroup genes with similar behaviours and unravel new regulation patterns [27, 47]. Alternatively, principal component analysis (PCA) can also be performed to reduce the dimension of the data set and classify genes according to their coordinates in a simpler system retaining the characteristic variability of the original data set [48]. Afterwards, presumed candidates highlighted from transcriptome analysis must be further validated both technically by qRT-PCR and functionally using reverse genetics, i.e. observing the effects of targeted gene deletion, overexpression or point mutation on the final phenotype [49-51].

Despite being a very powerful technology, microarrays, just like other hybridisation techniques, presents some drawbacks and do not capture the entire complexity of the transcriptome. First, DNA probes may be subject to cross-hybridisation with transcripts presenting sequences similar to the targeted one, thus affecting signal reliability [52]. Second, the abundance measurement is relative and its dynamic range is inherently limited upwards by signal saturation and downwards by background noise, reaching at most a hundred fold [53].

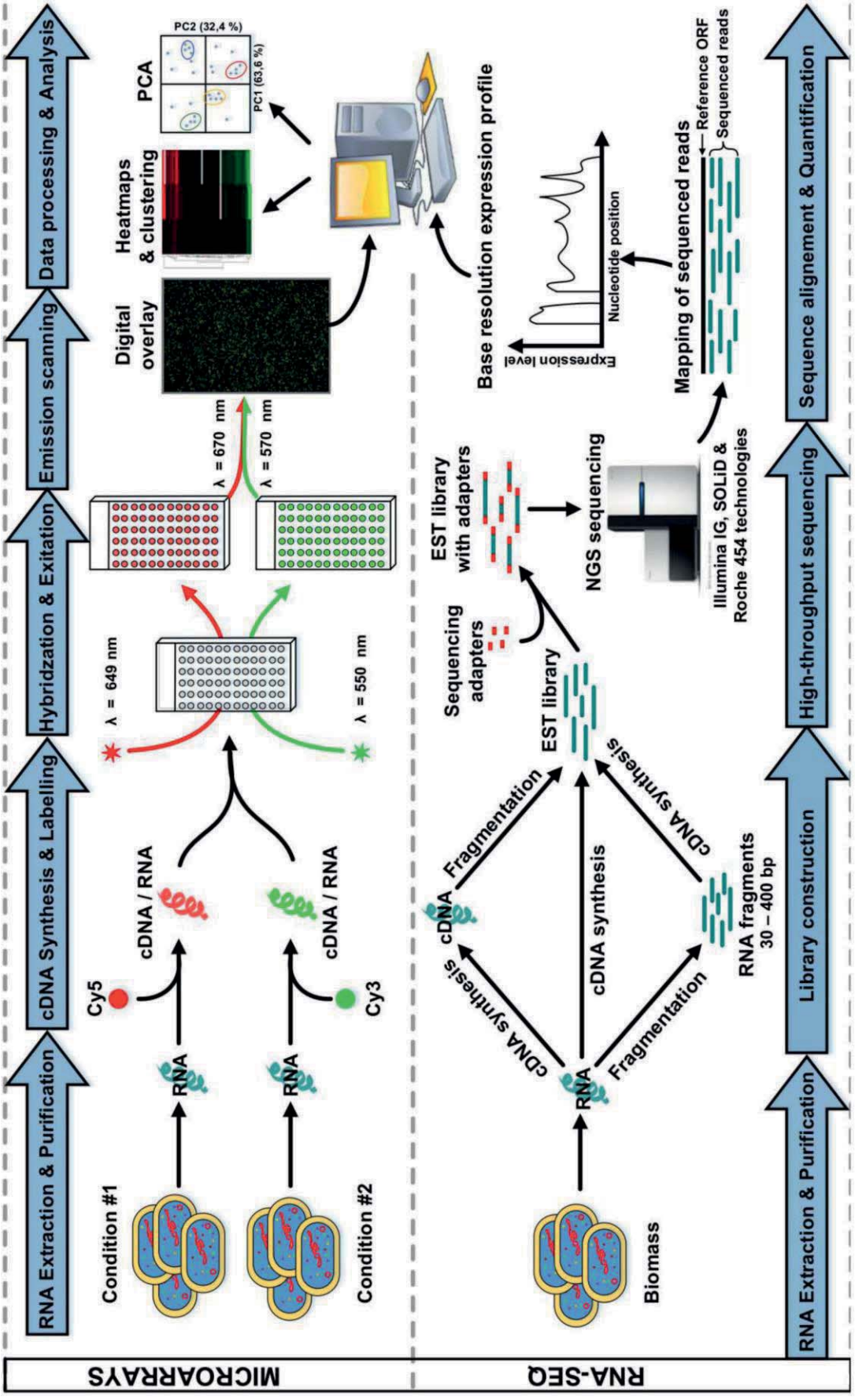

Figure 2.2: Analytical workflows for microarray- and RNA-seq-based transcriptome analysis – Central steps in microarray analysis include labelling of RNA extracted from two samples or of the corresponding cDNA with different cyanides, mixing and hybridisation of labelled samples on a designed array, excitation at dye-specific wavelengths and, finally, scanning of the array (upper part). For RNA-seq, an EST library is constructed by fragmentation and absolute transcript quantification is performed by high-throughput sequencing using next-generation sequencers (lower part). Regardless of the applied technique, generated data are post-processed using statistical methods such as analysis of variance, setting of cut-off values, principal component analysis (PCA) or hierarchical clustering to identify significant regulation. **cDNA**: complementary DNA; **Cy**: Cyanide; **EST**: expression sequence tag; **NGS**: Next-generation sequencing; **PCA**: principal component analysis. Figure was adapted from [53].

Furthermore, this technology relies on the knowledge of the genome under investigation and is therefore not generally applicable to non-model organisms. Even though many efforts have been devoted to increase the number and specificity of DNA probes, account for cross-hybridisation via mismatch probes and correction models or resolve labelling effects, saturation problems and alternative intramolecular folding, the future of transcriptome analysis might be somewhere else [54-58].

In fact, RNA-seq, a recently developed high-throughput technology based on next generation sequencing techniques, overcomes most of these limitations and is predicted to outperform microarray technology in the coming years [53, 59, 60]. To put it briefly, RNA samples are first cleared from abundant interfering ribosomal RNA, converted into their double-stranded complementary DNA (cDNA) and subsequently fragmented into small reads (30-400 bp) with DNase I [61]. Finally, those reads are ligatured with amplification adapters and massively parallel sequenced for absolute quantification and identification through mapping onto the reference genome if available (Fig. 2.2). There are ensuing benefits in terms of transcripts identification and quantification. First of all, the sequencing procedure enables the detection and characterisation of both known and unknown sequences with a single base accuracy and consequently single nucleotide polymorphisms as well as transcription boundaries and connections between exons can be resolved [53, 62]. Of particular interest is the possibility to study biological functions of intra- and intergenic non-coding RNA or particular transcription features such as directionality and allelic expression [61, 63]. From a technical point of view, this method is moreover less inclined to batch variation or background noise and the resulting reproducibility, sensitivity and dynamic range are therefore much greater than for microarrays, covering accurately expression levels up to 8000 fold [53, 64, 65]. Hence, RNA-seq is a very promising technology for uncovering complete transcriptomes but it currently still suffers a lack of hindsight compared to microarrays. So existing technical limitations or bias will probably only become clear as this technique spreads widely throughout scientific community.

2.1.3 Proteomics

Although transcriptome analysis gives a detailed and comprehensive overview of gene expression under given environmental conditions, detected mRNA transcripts are only intermediates between genes and proteins. On the contrary, proteins undertake the majority of cellular functions from catalysis to gene regulation, including nucleotides and amino acid recycling, signal transduction and structure stabilisation. Because of post-transcriptional regulations and protease activity, their concentrations can hardly be inferred from their transcript levels and must be assessed directly using dedicated methods and equipment [66, 67].

Proteomics deals with this specific issue and aims first and foremost at developing new analytical and computational techniques to detect, identify and quantify the whole set of proteins in a given sample, namely its proteome [68]. However, the scope of proteomics is much wider and also

includes the identification of post-translational modifications (PTM) and the detailed characterisation of protein localisation, interactions and structures that are essential to fully comprehend their biological functions [69, 70]. Although the field is still developing quickly, well-established approaches using various separation and quantification techniques are presently available and have been recently reviewed in detail [71, 72]. Historically, proteins were first separated according to their molar mass (MM) and isoelectric point (pI) by 1D/2D-sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis, subsequently stained with varied dyes, quantified using digital imaging and finally identified by GC-MS (**Fig. 2.3**) [73, 74]. This classical workflow is still well-suited for differential proteomics, the comparison of two protein samples, in particular after the development of difference gel electrophoresis (DIGE). In this method, proteins from two samples to compare are separately stained with two distinct cyanine-based dyes, then mixed and separated on a single gel, overcoming thereby the problem of gel variability inherent to the comparison of classical SDS-polyacrylamide gels [75, 76]. Despite great improvements of gel resolution through optimisation of buffer systems and gel compositions, this approach only enables a coverage of up to 50 % of the whole proteome, thus remaining inappropriate for global proteomics [77]. Indeed, only the more abundant non-hydrophobic proteins can be properly extracted from gels, whereas those presenting low natural abundances (10^3 – 10^4) cannot even be detected [69, 78]. Moreover, proteins with extreme pI (> 11 or < 3) or MM (> 200 kDa or < 10 kDa) can hardly be separated and conversely other proteins produce multiple spots or trains because of PTMs, making the subsequent identification and quantification difficult, if not impossible. Lastly, involved staining dyes and solvents for solubilisation of membrane proteins are often incompatible with GC-MS-measurements [71]. For these reasons, the use of off-gel chromatographic separation techniques and MS-based quantification methods have grown in importance in modern proteomics, whereas gel electrophoresis is mainly applied as a pre-fractionation step to reduce the degree of complexity of protein or peptide solutions to analyse.

Since the creation of the first mass spectrometer by Aston in 1919, a lot of progress has been made and soft ionisation methods such as matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionization (ESI) have enabled the measurement of intact proteins and peptides [79, 80]. However, the direct analysis of undamaged proteins or so called “top-down” strategy still requires high experimental efforts and the measurement of their constitutive peptides, namely the “bottom-up” strategy, remains in practice the method of choice for protein identification [81]. Here, protein samples are first enzymatically digested with a sequence-specific endoprotease like trypsin and resulting peptides are separated and fragmented in various ways in mass spectrometers (selected (SRM) or multiple (MRM) reaction monitoring) (Fig. 2.3). Their characteristic MS-fragmentation patterns are then used to identify the corresponding proteins and their eventual PTMs by comparing with theoretical mass spectra stored in databases (Fig. 2.3). Hence, this approach requires both high resolution mass spectrometers capable of performing exact mass determination over a wide dynamic range and powerful computational tools able to reconstruct proteins from their basic peptides.

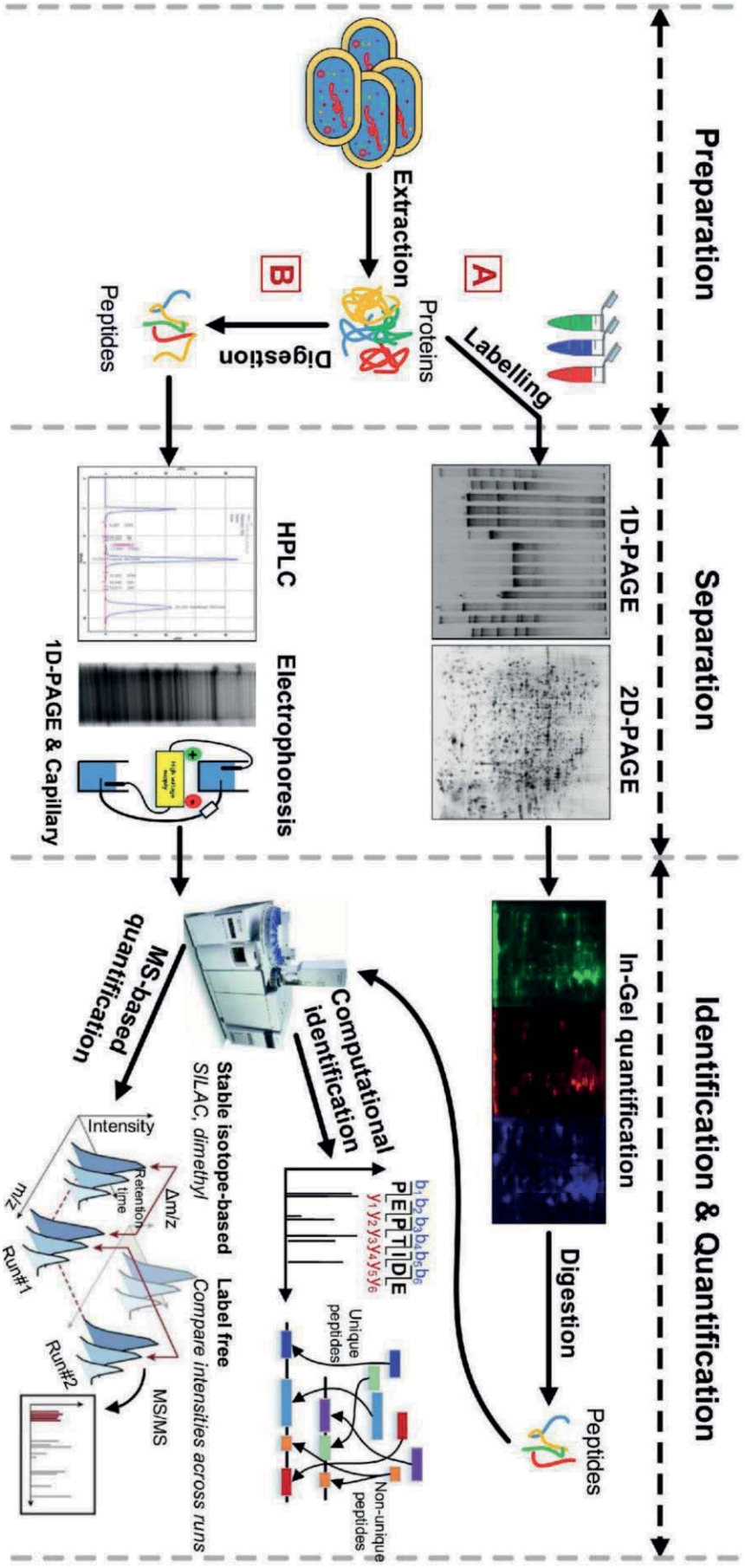

Figure 2.3: Analytical workflow for protein- (A) and peptide-based (B) proteomics – In the protein-based approach, extracted proteins are stained with different dyes for each sample, separated by gel electrophoresis and resulting colour intensities enable relative quantification. When used dyes are compatible, proteins can subsequently be extracted from gel, digested and submitted to GC-MS for identification. For peptide-based proteomics, proteins are first digested to peptides which are subsequently separated by HPLC or electrophoresis and finally identified and quantified by GC-MS using computational methods. Figure adapted from [1] and [3].

In addition to protein identification, the ongoing improvements of MS-proteomics and computational methods have now enabled their relative and absolute quantification on the basis of their peptide mass spectra. For relative quantification, the typical strategy relies on distinctive isotope tagging of proteins or peptides to compare. Indeed, as labelling does not affect physical properties, labelled and unlabelled peptides will be separated, ionised and fragmented in exactly the same way.
However, in the final MS-spectrum, the mass shift caused by the labelling will enable peptide differentiation. The labelling can be integrated directly into peptides or proteins by numerous chemical and enzymatic reactions (isobaric tag for relative and absolute quantitation (iTRAQ), isotope-coded affinity tag (ICAT), isotope-coded protein label (ICPL), enzyme mediated oxygen substitution (EMOS), acid mediated oxygen substitution (AMOS)) or, alternatively, it can be incorporated during growth on isotopically enriched medium (^{13}C, ^{15}N) or medium containing amino acid isotopes (stable isotope labelling by amino acids in cell culture (SILAC)) [82-84].
In most cases, this technique remains costly and label-free techniques based on algorithmic calculations have therefore gained interest in the past decades. They correlate protein quantity either with the intensity of mass spectra or with the number of peptides sequenced for a given protein (spectral counting). At the moment, these techniques are still limited in term of accuracy and mobilised great computational efforts. Nevertheless, with the development of effective algorithms to deconvolute and normalise MS spectra, they will undoubtedly become privileged methods in the future. Absolute quantification of proteins requires the use of internal standards (labelled or not) that are incorporated whether prior to or after protein digestion. Most of the time, the standard is a labelled version of the protein to quantify (protein standard for absolute quantification (PSAQ)) or a labelled peptide originating from this protein (absolute quantification of proteins (AQUA)) [85, 86]. Since the chemical synthesis of labelled proteins or peptides is very expensive, these techniques are often restricted to a small number of proteins in the framework of a targeted proteome analysis. To overcome this limitation, the QconCAT approach design a chimeric gene encoding selected signature peptides of all proteins to quantify and concatenating them into an artificial labelled protein. The purified chimeric protein is finally added to samples and enzymatic digestion generates automatically the labelled standard peptides necessary for absolute quantification [87-89].

2.1.4 Metabolomics

Since their introduction, genomic, transcriptomic and proteomic technologies have been successfully associated to gain new insights into the functional behaviour of biological systems [90-95]. This combination, however, has also rapidly started to show its limits and investigation of metabolites emerges as an essential counterpart to bridge the gap between genome and phenotype [96]. In fact, sequenced genome usually comprises 30-40% of genes encoding proteins with unknown functions or whose function was automatically attributed according to structural similarities, regardless of the potential biochemical significance of slight architectural differences [17, 97]. Moreover, whereas metabolite pools greatly depends on enzyme concentrations [98, 99], variations

in cell transcriptome and proteome do not necessary lead to altered phenotype, suggesting the existence of higher and post-translational regulation mechanisms [100-102]. As metabolites are further down the line from genome to phenotype and the connection nodes of all anabolic and catabolic reactions, their investigation arises quite naturally as the next step towards uncovering new gene functions, interactions, metabolic pathways and regulatory systems.

All metabolites synthesised by an organism under given physiological conditions constitute its metabolome [103]. Depending on the organism, it can encompass up to 200,000 metabolites varying significantly in their chemical nature and concentrations (from pM to mM) [104]. This diversity promises to be a very rich source of information but also makes the simultaneous identification and quantification of all metabolites, referred to as metabolomics, one of the biggest challenges of modern biochemistry. Indeed, no adequate measurement and sampling procedure have been developed so far to adequately recover and quantify the whole metabolome. Instead, modern techniques combining separation by gas (GC) or liquid chromatography (LC) with detection using mass spectrometry (MS), nuclear magnetic resonance (NMR) or infrared spectrometry (IR) have been employed for specific purposes, namely metabolite fingerprinting, target analysis and profiling (**Fig. 2.4**) [105, 106].

Metabolite fingerprinting aims at clustering different samples without quantifying, identifying or even separating metabolites, but only by using their characteristic measurement spectrum as discriminatory criterion. In clinical diagnosis, it is a systematic method for processing many samples and rapidly differentiating between healthy and diseased patient afterwards [107, 108]. Metabolite target analysis, for its part, is restricted to a small group of known compounds related to a given gene or specifically affected by a given abiotic perturbation. For this approach, metabolites of interest are extracted from samples using highly selective preparation and separation techniques.

Finally, metabolic profiling intends to identify and quantify different sets of defined metabolites such as amino acids, carbohydrates or those involved in a specific pathway in order to apprehend its function. This approach is often applied in pharmacology to trace the fate of administrated drugs and understand their effects. Thus, the current techniques are either too selective or not specific enough to reach a temporal separation of all metabolites. To extend the number of metabolites detected, composite metabolite profiling, a new approach involving simultaneous measurement of sample fractions with different systems, has been introduced. However, the additional spatial separation comes at a cost and the impact of other critical issues such as sample storage, measurement drift, matrix effects, sampling procedure and metabolite extraction on the subsequent quantification remains furthermore uncharacterised, underlining the need for suitable data normalisation methods [109]. The scope of metabolomics is huge and goes far beyond the single understanding of life. Indeed, unravelling functions of orphan genes or understanding interactions between metabolites and other biological components would for sure reveal new therapeutic targets and promising drugs. Moreover, the pharmaceutical industry is always on the lookout for new biomarker metabolites that make the spotting of health conditions easier. In addition, in the food industry, there is a growing interest for the discovery of new bioactive molecules and their incorporation in our everyday diet for promoting health and preventing diseases (functional food).

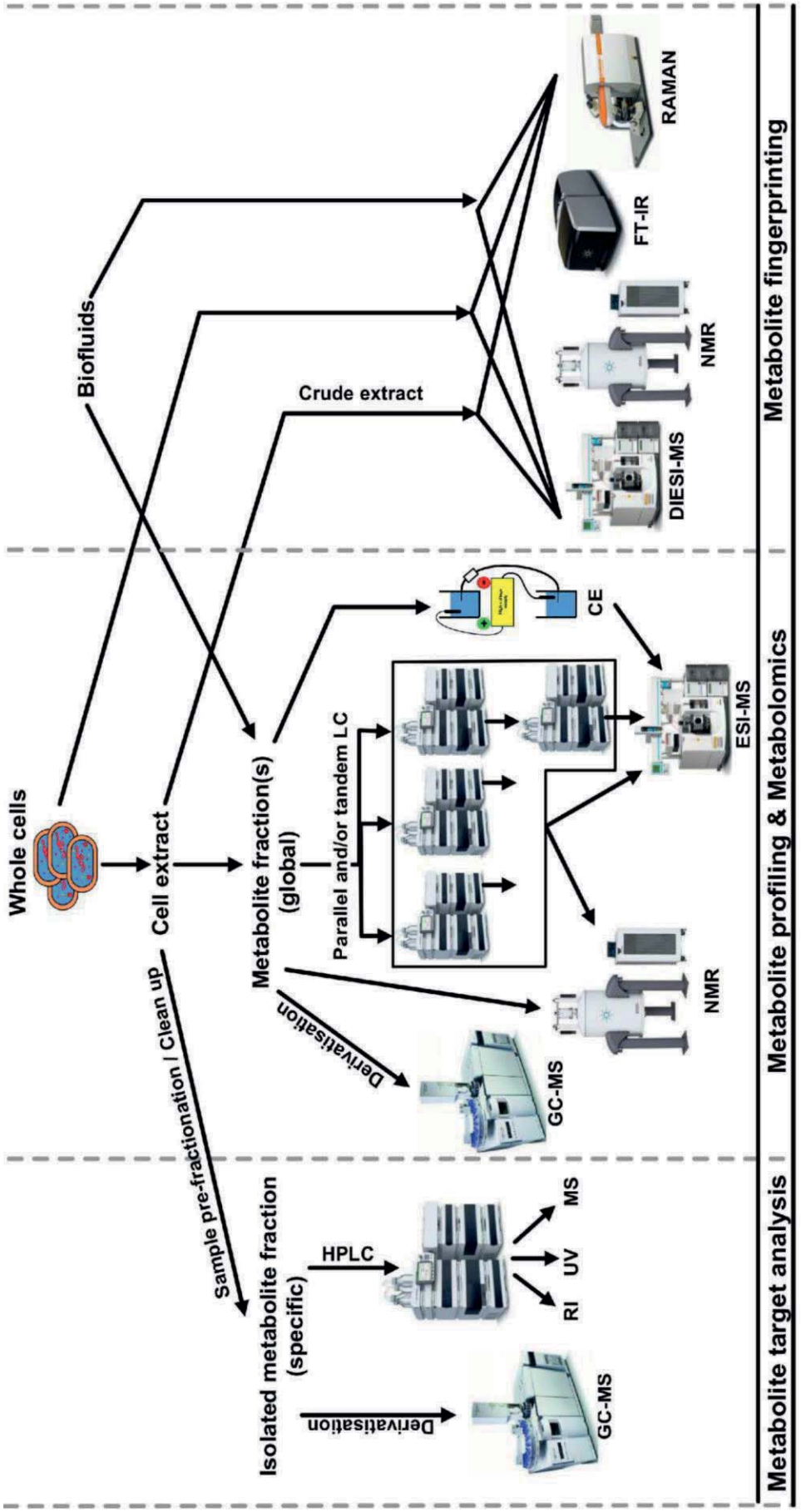

Figure 2.4: General strategies for metabolome analysis – Depending on the intended goal different approaches and equipments are used. Metabolite target analysis is focused on the quantification of a small and very specific group of known metabolites. Metabolite profiling aims at identifying and quantifying different sets of metabolites to unravel their metabolic function. Finally, metabolite fingerprinting only aims at clustering samples without identifying or quantifying metabolites by finding characteristic features in their measurement spectra. **CE**: capillary electrophoresis, **DIESI**: direct-infusion electron spray ionisation, **ESI**: electron spray ionisation, **FT-IR**: Fourrier transform infrared spectrometry, **GC**: Gas chromatography, **HPLC**: High-performance liquid chromatography **MS**: mass spectrometry **NMR**: nuclear magnetic resonance **RAMAN**: Raman spectroscopy **RI**: refraction index detection, **UV**: ultraviolet detection. Figure adapted from [110].