Toward genome-scale models of the Chinese hamster ovary cells: incentives, status and perspectives

Bioprocessing of the important Chinese hamster ovary (CHO) cell lines used for the production of biopharmaceuticals stands at the brink of several redefining events. In 2011, the field entered the genomics era, which has accelerated omics-based phenotyping of the cell lines. In this review we describe one possible application of this data: the generation of computational models for predictive and descriptive analysis of CHO cellular metabolism. We describe relevant advances in other organisms and how they can be applied to CHO cells. The immediate implications of the implementation of these methods will be accelerated development of the next generation of CHO cell lines and derived biopharmaceuticals.

It is generally appreciated that cell culture based on Chinese hamster ovary (CHO) cells holds substantial economical and medical importance. The global market for biologics was US$99 billion in 2009, where 60–70% of the products were produced in CHO cells [1]. Over 40 biopharmaceuticals have been produced in CHO cells so far, including monoclonal antibodies, hormones, cytokines and blood-coagulation factors. It is furthermore evident that the impact of CHO cell culture will only increase in the immediate future: the US market for biologics alone has been climbing from US$51.3 billion in 2010 to US$63.6 billion in 2012, and expected to increase at higher rates with the US Affordable Care Act [2]. The global market for biologics is expected to rise to US$190 billion in 2015 [3], and the percentage of CHO-derived products in approved new biologics are climbing. In 2010 and 2011 combined, 14 out of 19 approved biopharmaceuticals were derived from cell culture, the majority of these using CHO cells as hosts [4].

Despite this impact, the development of CHO cell processes – although highly successful – has been mainly driven by medium development and process engineering and to a lesser extent genomic technologies such as enhanced expression technologies for heterologous proteins [5]. Metabolic engineering, such as seen in microbial cell factories [6,7], has been very limited, although with some notable exceptions (for example, see [8,9]). We will argue that this has been due to the relatively late arrival of genome sequences for CHO cell lines; even though the first CHO expressed sequence tags (EST) sequences were published in 2005 [10], the first CHO genome sequences were published in 2011 [11,12], an entire decade after the first draft publication of the human genome [13], and two decades after the genome of the first eukaryote, Saccharomyces cerevisiae [14]. As a result, most early genome-based studies of CHO cells were performed by using the genome sequences from other mammals, for example, human, mouse or rat [15,16], which generally limits the possible experiments and interpretation of the results.

However, there is now ample genomic information available for the CHO cell lines. The CHO-K1 genome sequence [11] has been supplemented by the 2013 release of two draft genomes for the Chinese hamster (Cricetulus griseus) [17,18] from which the CHO cell line was originally isolated in 1957 [19]. Additionally, draft sequences for a number of CHO cell lines including the industrially relevant CHO-S and CHO DG44 have
been published [27]. While this still leaves a few widely used cell lines, for example, the CHO DXB11 cell line [20,21], unsequenced, and a general need for improved genome quality, it is clear that CHO cell research has reached the genomic era.

One highly promising application of genomics-based research is the generation of genome-scale models of CHO cells (Figure 1).

Potential applications of metabolic models to CHO cell cultures

A genome-scale metabolic model (GSM) is a systematic correlation of the genomic information of an organism to a metabolic network, effectively reconstructing the metabolic network of the cell type in question. Such a network is most often built from available generic pathway databases (e.g., Kyoto Encyclopedia of Genes and Genomes [KEGG] [22]) and specific literature for the organism being modeled, combined with an annotated genome [23]. This underlying network is often called a genome-scale reconstruction or genome-scale metabolic network reconstruction (GENRE). Integration of the GENRE with a linear programming-based mathematical framework allows modeling of the metabolic fluxes of the cell, which is often predictive and nearly always helpful in data interpretation. The actual model and computational framework apply the laws of mass conservation and balances of metabolic fluxes around single metabolites to compute enzymatic rates for every single enzyme present in the model. These rates are seen as averages for the culture and are most often given as specific rates relative to a certain number of cells. Additional algorithms may be applied to predict the effect of, for example, gene deletions/insertions, perturbations of feeding rates/nutrient uptake or increased production rates [24]. Pioneering work and additional application such as integration of the protein secretion network and regulatory information has been driven forward in *Escherichia coli* [25–27]. As CHO cells are arguably more complex in terms of gene numbers and cellular compartments than *E. coli*, the work associated with building a CHO GSM is more laborious and complicated, in particular in terms of assigning correct genes to enzymatic functions, and assigning enzymatic reactions to the correct compartments. However, the algorithms and uses of these models are general, and examples of potential applications from *E. coli* are equally relevant for CHO cells. In addition to this, implementation has been performed in a wide span of eukaryotic organisms as well, several of which with a complexity and quality of annotation resembling CHO cells. Examples of eukaryotic models include eukaryotic microbes, for example, industrially relevant yeasts such as *S. cerevisiae*, *Kluyveromyces lactis* and *Pichia pastoris* [28–31], filamentous fungi applied for enzyme production, for example, *Aspergillus niger* [32–34], and also higher eukaryotes such as *Arabidopsis* [35], mouse hybridoma cells [36], and human cells [37,38]. In these examples, cells, arguably as complex as CHO cells, have had their metabolism reconstructed. Cells from mouse, *Arabidopsis* and human are evidently of similar or higher complexity than the CHO cell. Even eukaryotic microbes such as filamentous fungi have a more complex growth physiology, with multicellular growth compared with the relatively homogeneous CHO cells with a more uniform growth. While the current annotation of the CHO genome is far from the quality of annotation and gene characterization found for human cells or even mouse cells [39], models can to a large part be generated by inferring function by homology to organisms with better annotation, for example, mouse or human in the case of CHO.

The primary applications of these models can be divided into at least five major categories: metabolic engineering, model-directed discovery, interpretations of phenotypes, analysis of network properties and studies of evolutionary processes [40,41]. All of these applications are highly relevant and interesting for CHO cell culture in their omics-driven approach to cellular physiology (Figure 1E).

Metabolic engineering holds considerable promise for CHO cell culture, as GSMS have the possibility of predicting the effect of gene deletions, additions and over-/under-expression. Several phenotypic traits of the CHO cells are sub-optimal for prolonged culture and protein productivity. Some examples of this are the conversion of high glycotic flux to lactate, or the formation of ammonium by conversion of amino acids in the medium. Both are detrimental to cell growth and product quality [42,43]. Accordingly, these processes have been subjected to metabolic...
Glycolytic flux: In this article, glycolytic flux is defined as being the conversion rate of glucose through glycolysis.

CHO genomics: Sequencing of multiple CHO cell lines and the progenitor hamster has revealed that while individual cell line genomes have a similar number of genes as the hamster, there is a large number of structural variations, in particular insertions, deletions and single-nucleotide polymorphisms. Such variation suggests that models should be tailored to individual cell lines.

**Key terms**

**Glycolytic flux:** In this article, glycolytic flux is defined as being the conversion rate of glucose through glycolysis.

**CHO genomics:** Sequencing of multiple CHO cell lines and the progenitor hamster has revealed that while individual cell line genomes have a similar number of genes as the hamster, there is a large number of structural variations, in particular insertions, deletions and single-nucleotide polymorphisms. Such variation suggests that models should be tailored to individual cell lines.
Review  Kaas, Fan, Weilguny, Kristensen, Kildegaard & Andersen

Model applications

Metabolic engineering
Model-driven discovery
Study of evolutionary processes

Interpretation of phenotypes
Analysis of network properties

Draft reconstruction of the metabolic network

Cell line-specific genomes, transcriptomes, proteomes, among others

Genomics
Transcriptomics
Proteomics

Expert knowledge
Databases

Experimental data

Iterative improvements of the model

Functional generalized mathematical model

\[ S \cdot v = 0 \]

\[ \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & ... & 1 \\
0 & 1 & 0 & 0 & 0 & ... & 0 \\
0 & 0 & 1 & 0 & 0 & ... & 1
\end{bmatrix} \]

\[ \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & ... & 1 \\
0 & 1 & 0 & 0 & 0 & ... & 0 \\
0 & 0 & 1 & 0 & 0 & ... & 1
\end{bmatrix} \]

Databases

CHO-S
CHO-GS
CHO-DXB11

CHO-K1

Model applications

Metabolic engineering
Model-driven discovery
Interpretation of phenotypes
Analysis of network properties

Future science group
the smallest possible set of essential metabolic genes in CHO cells, such as that achieved for individual pathways in *E. coli* [55] or *S. cerevisiae* [56]. Such identification could be applied for generating cell lines with a trimmed metabolism, thus decreasing the variability of the system. This would be feasible for CHO, despite the presence of extra isoenzymes or alternate pathways in many enzymatic steps. For several enzymatic functions, only one of the isoenzymes is detected at the protein level [57]. Alternatively, one can delete steps in the enzymatic pathways/elementary flux modes, where only one enzyme exists.

Studies of evolutionary processes have been a recurring theme in several applications of bacterial network, in particular the *E. coli* GENRE [40,41]. In such studies, specialized models have, for example, been developed to describe specific strains of *E. coli*, and compared these to identify the genetic origin of specific phenotypes [58]. Given the availability of the genomic sequence for multiple CHO cell lines with varying properties [17] and surely more to come, such an exercise would hold exciting perspectives for interpreting these genomes. One possible application would be the genetic basis for certain metabolic features in cell lines generated by mutagenesis, and the possibility of *de novo* engineering the features into a ‘clean’ background. This would reduce possible complications from decreased genetic stability in mutants subjected to mutagenesis [59].

With the above-mentioned being only a small percentage of the possible applications of such models for CHO cell lines, the potential is clearly large for the generation and application of GENREs and GSMs for CHO cells.

**Additional available data sources for increased applicability of CHO models**

The availability of an annotated genome for CHO cells is the bare minimum of information required to generate a draft model for CHO cells. Models of microbial systems have been published based mainly on literature on characterized genes (e.g., for *Corynebacterium glutamicum* [60] or *Aspergillus niger* [32,41]), but this requires detailed legacy data for a wide selection of metabolic pathways. In general, most recent generation of GSMs is based on variations of a standardized protocol for metabolic network reconstruction and model validation published in 2010 [23], using basic genome annotation as a starting point for the organism of choice.

However, other types of omics data have proven highly valuable for model generation, validation and application. Here, the CHO field is maturing at an impressive pace, considering the quite recent publication of the first public CHO genome [11], followed by a wealth of other omics types being published in these years [62]. Here we will briefly emphasize selected studies which provide data highly applicable to CHO modeling, either due to the experimental setup of the study, the type of the data or the perspectives these offer for CHO modeling.

**Genomics**

A well-annotated genome with a high coverage is a crucial component in building a GSM with predictive power or a GENRE with a potential for informative data integration and interpretation. It is essential to be able to identify the genes of all major metabolic pathways in order to generate an accurate GENRE and following that a GSM. This requires a genome coverage of ideally >99% of the genes. Early EST sequencing efforts [10] identified only less than 20% of the genes, shown to be present in the CHO-K1 cell line draft genome [11]. Convenient for model construction, the genome sequence has been made accessible at the online database CHOgenomic.org [63,64] as well as at the NCBI genbank. The coverage appears to be at least 99%, at least it was demonstrated that homologs existed for 99% of the genes in the human genome associated with glycosylation.

Due to the variability of the cell lines, it can be argued that it would be most appropriate to use the progenitor Chinese hamster (*C. griseus*) both as the source of a reference genome and as a scaffold for a master CHO GSM, from which specialized models can be generated for individual cell lines. This is now possible due to
the publication of draft genomes for \textit{C. griseus} \cite{17,18}. This publication showed that there are more than 3.7 million point mutations between the progenitor hamster and the CHO cell lines \cite{17} and extensive chromosomal rearrangements that have occurred between CHO-K1 and CHO-DG44 \cite{65} emphasizing the effects of the mutagenesis that occurred in the process of creating the various cell lines. Currently, in order to fully exploit these sequences, these genomes must be mapped against reference genomes with gene annotation. Improving the gene annotation for the \textit{C. griseus} genome and CHO-K1, which has become the de facto reference genome for cell lines, would be beneficial to model building. Current and new genomes could thus be aligned to these references for detection of mutations.

Furthermore, the state of genome assembly should be improved. Currently the genomes for both hamster and cell lines are divided into at least 4000 contigs per genome, which means that genes for important metabolic functions may be lost in the sequencing gaps. Such gaps can to some extent be detected and fixed in the network reconstruction process \cite{23,32}. Even so, an appropriate solution would be to apply third generation sequencing to yield longer sequenced reads that can assemble the contigs to improve the coverage of the genomes. Such efforts are in progress in the community \cite{Borth N, Pers. Comm.}, and should have a substantial positive impact on the models, which can be constructed for CHO cells.

\textbf{Transcriptomics}

In general, it is only a low percentage of the CHO genes which are actually expressed under normal condition, for example, only approximately 50% of the genes involved in protein glycosylation are transcriptionally active \cite{11}. Consequently, integration of dynamic omics data such as transcriptomics and proteomics is important for accurate prediction of gene deletion/silencing effects. Several studies and tools are now available for this, including both sequencing and DNA microarray based methods. Naturally, some of the first transcriptome data were generated prior to the genome sequence based on EST sequences from CHO and mouse used for design of microarrays \cite{66,67}.

Worth particular mention is a large-scale comparison of microarray data from more than 120 individual CHO cultures \cite{68}. The data can be accessed through the web-based CHO gene coexpression database allowing easy access to the list of genes found to coexpress with, for example, cell specific productivity and growth rate. Such data could be used for model improvement and validation. For easy and relatively inexpensive assessment of the CHO transcriptome in future experiments, a new generation of the Affymetrix\textsuperscript{®} CHO DNA microarray (Affymetrix, CA, USA) has been launched with up to 26 unique sequences of each transcript with a total of more than 644,000 probes \cite{69}.

RNA-sequencing is expanding for CHO culture as in many other fields \cite{70}. Recently, a transcriptome database for CHO RNA sequencing data has been developed and is available at GenDBE \cite{71,72}.

In summary, transcriptomics data are abundantly available, and will only increase in the coming years.

\textbf{Proteomics}

The CHO proteome is interesting in the context of CHO metabolic modeling as it can provide additional functional information, in some cases expanding on transcriptomic evidence. The proteome of CHO-K1 was thoroughly characterized by Baycin-Hizal in 2012 \cite{57}. Here, 6164 proteins were detected. Of these, only 60% were also detected at the mRNA level by Xu in 2011 \cite{11}. The functional application of the data and the need for having models specialized to individual cell lines become apparent from this study. Statistical analysis indicated that some pathways such as fatty acid metabolism, amino sugar and nucleotide sugar metabolism, which provide important precursors for recombinant protein synthesis, as well as protein processing and apoptosis, were enriched in CHO-K1 \cite{57}.

Given the principal application of CHO cells for production of secreted proteins, in this context, secretome data, such as characterized from the CHO DG44 and CHO-S cell lines by Slade \cite{73}, are interesting to incorporate. Such data can help identify secretory bottlenecks or extracellular proteases as seen for microbial cell factories \cite{74,75}.

CHO-specific protein databases have been constructed based on data from the CHO-K1 genome \cite{11} and the CHO transcriptome \cite{76}, and have been shown to increase the number of identified proteins by 40–50% from proteomics studies compared with only using protein databases based on, for example, the murine proteome \cite{77}.

It is generally accepted that the generation of proteomics data is more technically challenging than transcriptome data, but the pilot studies within CHO cells, such as those mentioned above, show that there is clearly additional value to be gained from interrogating this data set.

\textbf{Metabolomics}

As mentioned in the text above, metabolomics have considerable value to add in the model building pro-
cess, as this type of data can help identify metabolic pathways, which are experimentally shown to be occurring in the cells due to the presence of metabolic intermediates or products, but the genetic basis is not necessarily clear. Being able to identify and include these pathways can increase the predictive power of the model.

For such inclusion, several studies of high-quality and standardized protocols [78,79] have been available for several years, as metabolomics are not as such dependent on the availability of a genome.

Two studies are of particular interest in terms of getting data of a sufficient quality for model integration. The first was published by Dietmair et al. [83] correlating intracellular and extracellular metabolite concentrations with growth. The second is the work of Chong et al. [84] where intracellular metabolite profiles were obtained for eight single-cell clones with high and low production rates of monoclonal antibodies at the mid-exponential phase during shake flask batch cultures. Such studies can give insight in metabolic responses, and help validate CHO metabolic models, in that one can examine and adapt the ability of the model to predict these responses.

The application of the metabolic networks to interpret metabolomics data can also be exemplified in a 2012 study by Selvarasu et al. [85], where a generalized metabolic network of mammalian cells was adapted to CHO cells to aid in metabolomics data interpretation (see further details below). The coupling of the network, genome-scale-modeling and metabolomics data allowed the identification of growth-limiting factors.

Overall, the CHO field is at this point uniquely poised to utilize the substantial amounts of available omics data in building high-quality models for CHO cells. An overview is presented in Figure 1. During any future model-building efforts, one should draw upon the current availability of computational models for CHO and similar systems, and incorporate this where appropriate.

**Overview of cellular modeling efforts in CHO cells & beyond**

So far, no dedicated effort to building a CHO GSM de novo has been published. The closest example is the adaptation of a model of mouse hybridoma cells [36] to CHO cells by the addition of 35 CHO-specific metabolic reactions and subsequent model curation resulting in a model comprising 1540 reactions and 1302 metabolites [85]. This model has been further developed by other groups, although not published through a journal at this time, but is available for download from CHO.sf.net [86]. A similar approach of adapting a mouse GSM was employed by Martínez et al. [87] for examining the energy consumption and metabolism surrounding lactate formation and consumption in CHO cells.

Dedicated models have been developed for related cell lines in other systems, as mentioned a generic model for mouse cells, applied to mouse hybridoma cell lines [36], and a model for the HEK-293 cell line has been developed as well [88]. This study is particularly promising for CHO cell modeling, as the HEK-293 model was developed by reducing the generic model for human cellular metabolism [37] to a model specific for HEK-293 metabolism. Furthermore, this model was employed to interpret both transcriptomic, metabolomic and flux data to gain functional understanding of glucose and glutamine metabolism; both key features for CHO metabolism [88]. A similar study has been seen for baby hamster kidney (BHK) cells for interpretation of metabolomics data [89].

These models listed above represent the full list of available metabolic genome-scale models with relevance to CHO cells. However, to the best of our knowledge, a model specific for CHO cells or any specific cell line has still not been generated.

One area, where modeling in CHO cells is more developed, is the kinetic modeling of protein N-glycosylation, in particular integrated with mass spectrometry on glycans. Here, very accurate predictions and substantial networks have been generated and improved over the last two decades. The first mathematical model for protein N-glycosylation process was built in 1997 by the complementary studies of a single-compartmental model [90] and a multi-compartmental model [91]. Later work expanded upon the previous work to involve glycosylation processes as galactosylation, fucosylation, sialylation and addition of N-acetyllactosamine residues [92]. This model had up to 7565 N-glycans and 22,871 reactions included. Furthermore, two glycosylation models based on different views of protein transport across the Golgi, namely Golgi maturation mechanism and vesicular transport mechanism, were studied and compared. This model was highly expanded and sophisticated by the same group to include interpretative power of N-glycan mass spectrometry data [93]. More recently, an optimized model considering 77 N-glycans, 8 enzymes, 4 nucleotide transporters and 95 reactions with individual rate expressions were built on the basis of Golgi maturation mechanism with an improvement of taking Golgi protein recycling into account [94]. On top of that, a more comprehensive glycosylation model that links a model that described the metabolism of nucleotides and nucleotide sugars...
to the previous N-glycosylation model was developed by the same group [95]. These networks have been shown to have both high predictive and interpretative power, and would be unique key features to have integrated in CHO GSMs, to the extent that it is possible. Such additions could predict effects of glycosylation engineering and/or the effect of different substrate uptake rates.

**Conclusion & future perspective**

With the potential of GSMs tailored to CHO cells as demonstrated above, it is not surprising that several groups in the CHO community are working on building whole and partial reconstructions of CHO metabolism, including some of the authors of this review. These groups have this year formed a consortium compiling their work, and are working toward generating a community consensus model for CHO cells [Lewis NE, Pers. Comm.]. Such models and network reconstructions are known from several other research communities, including Salmonella Typhimurium [96], yeast [29,97] and human [37] metabolism.

The future arrival of the CHO GSM will probably raise the same discussion that followed the release of the first CHO genome: How well does this model describe each of the different CHO cell lines? Each of the cell lines has undergone rearrangements and has diverse transcriptomes and for this reason several parameters will need to be investigated. Future and current sequencing projects for individual cell lines should be combined with bioreactor characterization of the cell lines and their corresponding models to gain a functional understanding of the differences (Figure 1C–D). The next years will tell whether these models will be able to model the complex behavior of the CHO cell and open up new design targets such as it has been the case in microbes. Making such specialized models will be a substantial amount of work, but this task will be made easier, if a generic CHO model of high quality based on an assembled and annotated reference genome is generated first. From that, specialized models can be made in semi-automated fashion through comparative genomics. This would have the additional advantage that annotation of the genomes of the individual cell lines would not be required (as this is currently not available [17]), but could be achieved by alignment to the reference genome.

Should the models be able to deliver on the promise and potential seen in other cells, it is bound to trigger a second wave of CHO cell line engineering. Notably CRISPR-Cas9-based genome-editing systems being made available at non-cost prohibitive prices [98] and efficient high-throughput mammalian vector design systems [99] support the development of faster and cheaper genome engineering tools to accelerate future cell line engineering efforts in CHO.

In summary, the potential of genome-scale models stands to be unleashed in CHO cells within a very short time span. Combined with the genomes ushering in the genomics era for CHO, substantial amounts of omics data are being generated, and the development of efficient genetic engineering tools of CHO cell culture will soon move into the next generation of cell line development. Such advances promise better and cheaper development of biopharmaceuticals for this important group of cell factories.

**Financial & competing interests disclosure**

Y Fann and D Weilguny are employees of Symphogen A/S, CS Kaas and C Kristensen are employees of Novo Nordisk A/S. While the companies have funded the salaries of the employees, the company has had no role in defining the content of the manuscript. MR Andersen and HF Kildegaard declare no competing financial interests. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

---

**Executive summary**

- Genome-scale metabolic models have been applied with success in many other prokaryotic and eukaryotic cell factories.
- The Chinese hamster ovary (CHO) field now has all of the relevant information and methods needed to construct and apply such models.
- A CHO metabolic model will have applications both in design and engineering of cells, but equally important also in interpretation of omics data. The potential is large.
- Initial CHO models have adapted from models of mouse metabolism, but no de novo CHO models have been published at this time.
- The community is currently constructing a consensus model for CHO metabolism.
- Added value will come from generating specialized models for individual cell lines.
References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

• Article is a solid example of metabolic engineering of Chinese hamster ovary (CHO) cells.
** Expression of glutamine synthetase in CHO cells is an important technology as well as an example of a type of engineering which can be predicted using genome-scale metabolic models.
33 Vongsangnak W, Olsen P, Hansen K, Krogsgaard S, Nielsen J. Improved annotation through genome-scale


• Gagnon et al. demonstrates how process design in many cases can improve fed-batch performance.


50 Lewis NE, Hixson KK, Conrad TM et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. 6, 590 (2010).


•• Provides a solid overview of demonstrated data integration approaches for metabolic networks.


Towards genome-scale models of the Chinese hamster ovary cells

64 CHOgenome.org. www.chogenome.org/
72 GenDBE – ProCell. A eukaryotic genome browser and annotation system. https://gendbe.cebitec.uni-bielefeld.de/cho.html
86 Possibly the best example of the application of genome-scale metabolic modeling to CHO cells.
87 CHO model download. http://cho.sourceforge.net/
89 • Martínez et al. have a solid analysis of carbon fluxes through the CHO cell.
91 Vernardis SI, Goudar CT, Klapa MI. Metabolic profiling reveals that time related physiological changes in mammalian cell perfusion cultures are bioreactor scale independent. Metab. Eng. 19, 1–9 (2013).
96 Jiménez Del Val I, Nagy JM, Kontoravdi C. A dynamic mathematical model for monoclonal antibody N-linked


