Biosystems and Information Technology (2013) Vol.2(1) 1-5
DOI: http://dx.doi.org/10.11592/bit.130501

original scientific article

Automatic comparison of metabolites names: impact of criteria thresholds

Martins Mednis'[1*], Armands Vigants[2]

[1]*Biosystems Group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV-3001, Jelgava, Latvia*
[2]*Institute of Microbiology and Biotechnology, Faculty of Biology, Latvia University, Kronvalda bulv. 4, LV-1586, Riga, Latvia*
Corresponding author
martins.mednis@llu.lv

Received: 15 May 2013; accepted: 29 May 2013; published online: 29 May 2013.
This paper has supplementary material.

Abstract: *The growing number of stoichiometric reconstructions and models tends to change the model building process. Instead of creating a new model from scratch scientists can look at the earlier created relevant models to assess the opinion and consensus level of other modellers. Several initiatives have been performed to build consensus models for particular organisms following this approach. One of possible improvements in the model development taking into account earlier developed ones is automated comparison of models. That is enabled by the fact that models usually are in a computer readable format to be simulated. Still there are some problems like different ways of naming metabolites, models without formulae of metabolites, different approach in definition of compartments and other peculiarities of different research groups at different times.*

There are several software tools that offer reconciliation or mapping of metabolites in models to assess the similarity of models. It is computationally trivial to find metabolite pairs with identical names in two models. Still very often the comparison algorithms lack flexibility and some work should be done by manual curation of metabolite pairs to recognize that the difference is caused by symbols like brackets, quotes, apostrophes, spaces, upper/lower case letters or similar ones.

The proposed approach suggests combination of automated comparison with manual curation: most of possible metabolite pairs are rejected by computer leaving just the most similar metabolite pairs for manual comparison. The elasticity in metabolite name comparison is introduced using Levenstein similarity ratio and Levenstein edit distance. Application of these criteria with different acceptance thresholds is analyzed comparing two models of Saccaromyces cerevisiae with 681 and 1063 metabolites. The results are compared with manually approved pairs of matching metabolites.

Keywords: fuzzy string comparison, edit distance, similarity ratio, metabolic networks.

1. Introduction

The molecular processes in cells form a huge network, which makes detailed mathematical modeling and simulation extremely difficult (Schulz et al., 2006). Genome-scale reconstructions of metabolic networks and stoichiometric models may contain thousands of metabolites and reactions (Thiele et al., 2013). The functions of such networks are hard for the human mind to comprehend (Palsson, 2006). The process of iterative model building (Thiele and Palsson, 2010) promises to accelerate biological discovery, product development, and process design (Ideker et al., 2001; Palsson, 2006). The increasing knowledge base of living organisms leads to even more complex biochemical models and scientists often decide to model only a part of genome, not the whole metabolism (Mednis and Aurich, 2012).

It is wise to look around and check what other models of that particular organism exists before creating a new model of an organism or it's part. If more than one model is available, it is important to evaluate them and choose as starting point the most comprehensive model, the one with least inconsistencies, intersection or merge of models. The published genome-scale reconstructions of the same organism should be carefully compared to avoid misleading conclusions. Consequently, the need for analysis, comparison, intersection and merge of biomodels is growing. The demand for a method to relate different models has been pointed out (Gay et al., 2010; Radulescu et al., 2008).

Metabolites could be compared by chemical formula and name. Due to identical formulas in case of isomers (Poggendorff, 1830) and the lack of formula in many models the comparison of metabolite names becomes irreplaceable. Still chemical formulas also can serve as additional criterion during comparison of models. If the formulas are available, it is possible to check if they are equal for the particular pair of metabolites. Still the main criterion remains metabolite name.

Comparison is an essential procedure before merging or intersecting two or more models. Some model comparison related functionality is proposed by existing software tools. The COBRA toolbox (Schellenberger et al., 2011) has two functions related to the search for duplicates and the comparison of two models. *CheckCobraModelUnique()* finds reactions and metabolites that are not unique. The second function *isSameCobraModel()* receives two models as input parameters and returns three outputs: *isSame* - “true” if all common fields are identical, otherwise “false”; *nDiff* - number of differences between the two models for each field.

The FAME (*Flux Analysis and Modeling Environment)* (Boele et al., 2012), is a browser-based graphical interface that allows users to build, edit, run, and visualize stoichiometric models. The Fame also has comparison and merge functionality. The merge facility can be used to merge an additional pathway into an existing model. When merging, one model must be designated the *master* model, and the other the *slave model*. This is to ensure that whenever a merging conflict exists, e.g. when a reaction exists in both models but with different constraints, the information in the master model will take precedence.

If reactions have the same reaction ID in both models (this is usually the case for identical reactions if both models are generated by the same source, such as FAME), they are assumed identical will be included only once. However, FAME has no way of knowing whether reactions are identical if neither reaction IDs nor species IDs follow the same convention - if this is the case, both versions of the same reaction will be included. If a reaction has been deleted from the master model, but exists in the slave model, it will be in the result of the merge operation (existing reactions overwrite deleted ones, as no record of deletions is kept).

The above described tools do not tolerate even small differences in metabolite names like brackets, quotes, apostrophes, spaces, upper/lower case letters and some more symbols which may be caused by the modelers style of defining metabolites. Therefore many pairs of identical metabolites may not be recognized leading to wrong conclusions about the similarity of models. Some other tools compare models in a more flexible and adaptive way.

Model SBMLmerge feature is provided in SemanticSBML (Krause et al., 2010). SBMLmerge (Schulz et al., 2006) first merges the lists of elements in the annotated input files. The resulting list is then searched for conflicting elements by pairwise comparison, based on the identifying attributes, including the annotations. If two conflicting elements are found, their describing attributes are compared. The values for these attributes can be identical for both conflicting elements, or they can differ in one or more values: if all attribute values are identical, the elements are assumed to have the same biological meaning.

If both networks (models) come from the same source or both models contain information about metabolites chemical formula, it is possible to skip reconciliation of metabolites. An approach of reaction comparison based on comparison of metabolites formulas has been described by Mednis (Mednis et al., 2012; Mednis et al., 2012). However, such approach is not suitable when metabolite chemical formulas are not available.

A software tool ModeRator (Mednis et al., 2012) for biochemical network comparison can perform metabolite mapping between two models based on name similarity and other criteria as well. The software tool also introduces three-level-filtering algorithm (Mednis and Aurich, 2012) which significantly reduces the amount of data that requires manual curation.

This article is devoted to the efficiency analysis of additional comparison criteria in application of mapping metabolite names. The impact of filtering by similarity ratio and edit distance, metabolite compartments and Three-level-filtering is discussed. Two *Saccharomyces cerevisiae* (Baker's yeast) models are compared to determine how the thresholds of name similarity criteria impact the number of manually approvable pairs.

2. Materials and methods

2.1. Pairwise comparison

The purpose of metabolite *mapping* or *reconciliation* (Oberhardt et al., 2011) is to find a corresponding metabolite in Network B for each metabolite from Network A (Fig. 1).

Comparing two lists of metabolites can be very laborious since it involves the screening of all possible element combinations. Therefore software automatically rejects invalid combinations leaving only the most similar pairs of metabolites for manual curation.

The algorithm of metabolites reconciliation starts with creation of Cartesian product from both lists of metabolites. The result is another list containing all possible pairs of metabolites (Fig. 1). The next step is to calculate name similarities for all metabolite pairs in this list.

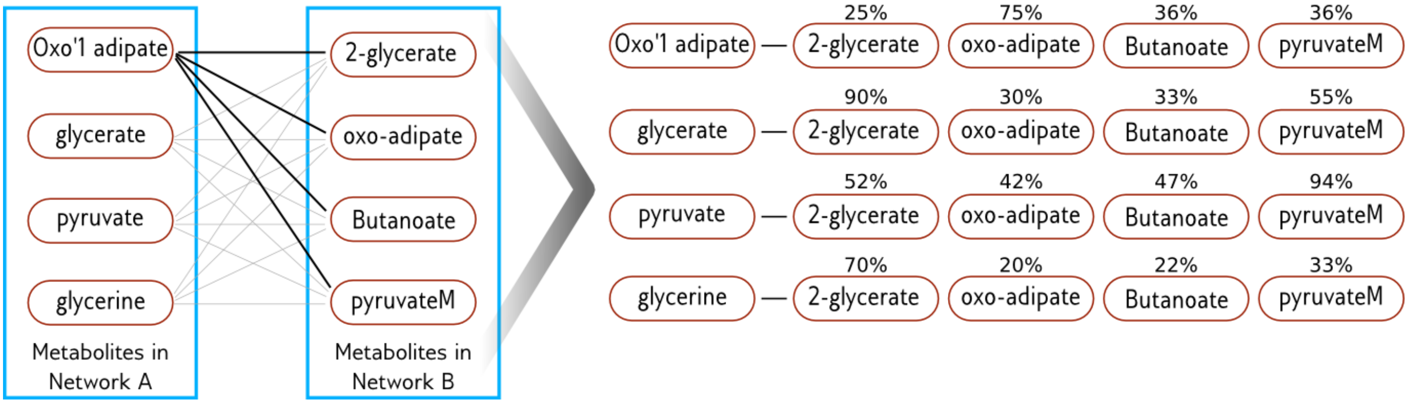

Fig. 1. **Pairwise comparison before Three-level-filtering.**

1st filter

The 1st filter is a set of user defined thresholds for various criteria. In the example discussed in this section, the criterion is name similarity ratio. We used Levenstein edit distance (Levenshtein, 1966) and similarity ratio implementation in pylevenstein (Mulligan, 2013). The threshold in this example is 55%. It means that each pair where metabolites names similarity ratio is below the threshold will be automatically discarded leaving only the most relevant results (Fig. 2). The edit distance can be used as a criterion. In such case, the algorithm automatically discards pairs with edit distance above the threshold.

The user can define also the compartment threshold. Filtering *by compartments* takes into account the information about metabolites compartments and will automatically discard those metabolite pairs where the information about their compartments is conflicting. In a real-world example a pair of *glucose[cytosol]* and *glucose[extracellular]* would be discarded, but a pair with *L-lysine[cytosol]* and *L-lysine[cytosol]* would be spared because the compartments are matching.

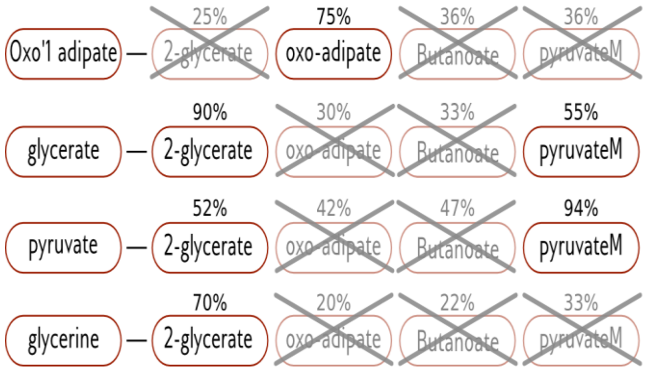

Fig. 2. **Three-level-filtering: 1st filfer.**

2nd filter

The 2nd filter discards pairs with least similarity. It is possible that many pairs with the same metabolite have similarity higher than the threshold. In Fig. 3 the metabolite “glycerate” has *connection* with metabolite “2-glycerate” and “pyruvateM”. However, the similarity with metabolite “2-glycerate” is higher and therefore all other *connections* to metabolite “glycerate” are discarded. The algorithm iterates through all *connections* of each metabolite from Network A and discards pairs with similarity less than highest.

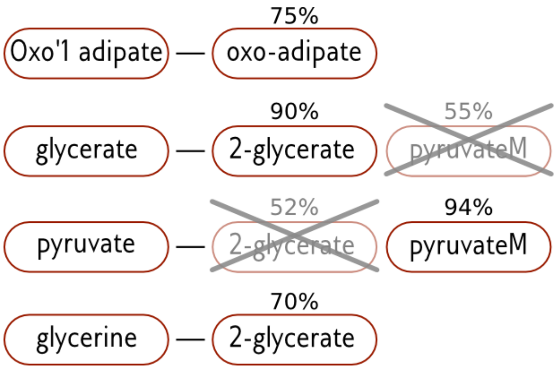

Fig. 3. **Three-level-filtering: 2nd filfer.**

3rd filter

The 3rd filter also discards pairs with least similarity. While the 2nd filter eliminates multiple connections to the same metabolite in Network A (Fig. 3), the 3rd filter does the same thing, but in opposite direction. It eliminates multiple connections to the same metabolite in Network B by discarding pairs with similarity ratio less than highest (Fig. 4).

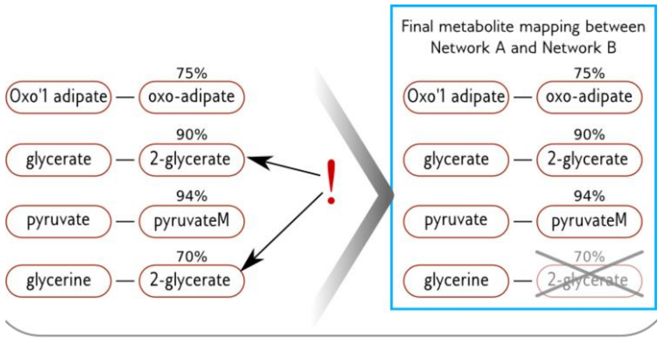

Fig. 4. **Three-level-filtering: 3rd filfer.**

2.2. Application of thresholds of different criteria

The threshold for similarity between two metabolite names is a number (set by the user) ranging from 0 to 100%. Examples of similar names and their similarity ratios are listed in Table 1.

Table 1.

Example of similarity ratio and edit distance variations for different strings

String A	String B	Similarity ratio	Edit distance
Bicarbonate	bicarbonate	0.9	1
Glucose-6-phosphate	Glucose six phosphate	0.8	5
Glucose-6-phosphate	Glucose-six-phosphate	0.9	3
L-tryptophanyl-tRNAtrp	L-Tryptophanyl-tRNA(trp)	0.86	4
L-lysine	L-Lysine	0.87	1
D-glutamate	L-Glutamate	0.81	2

While the *three-level-filtering* algorithm does not solve the problem of fully automatic comparison, it can drastically reduce the amount of data that requires manual approval (Mednis and Aurich, 2012).

3. Results and discussion

In the experiments described in this article two *Saccharomyces cerevisiae* (Baker's yeast) models were compared. The models having 904 and 1268 reactions are based on iND750 (Duarte et al., 2004) and iLL672 (Kuepfer et al., 2005). Both models have located metabolites and reactions in compartments. In all experiments similarity ratio threshold was decreased step-wise from 100% to 5% with the step size 5% (except where it is noted).

The total number of metabolite pairs to process is 723903 and is formed by a multiplication of metabolite numbers in both models – 1063 and 681. In this case the maximal possible number of valid pairs (MPNVP) can not exceed 681 as that is the number of metabolites in the smallest model. The effect of all the three filters at different similarity ratio thresholds (Fig. 5) indicate almost linear increase of filter 1 passing metabolite pairs. Threshold “0” means that 1st filter is not operating at all. Therefore the curve reaches the total number of metabolite pairs when threshold is low. The curves of 2nd and 3rd filter stop growing when the threshold is below 40%. Thus the 1st filter functions as a pre-filter and the realistic number of combinations for manual comparison is determined by 2nd and 3rd filter.

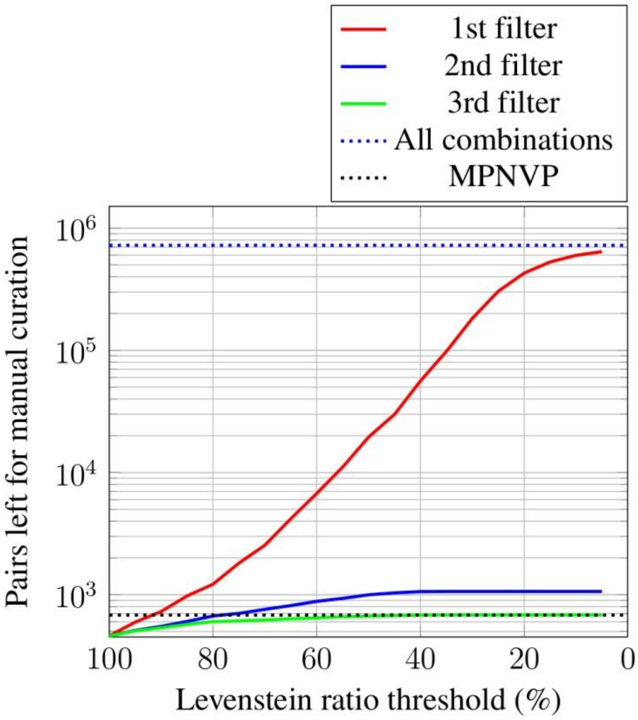

Fig. 5. **The effect of three-level-filtering**.

In the Fig. 5 the number of matched pairs after the 3rd filter at low similarity ratio threshold reaches MPNVP: maximal number of possible pairs. Introduction of compartment criterion reduces the number of mapped pairs to 60% of MPNVP (Fig. 6) demonstrating the high importance of correct compartment handling during comparison. Even at the similarity ratio threshold of 100% the number of metabolites with identical names reduces the number of matched pairs by half (48%) when compartments are taken into account.

Different criteria can be combined while comparing metabolite names. Additional introduction of edit distance to similarity ratio (compartments not taken into account) (Fig. 7) reduces the number of mapped pairs.

In the Fig. 7 legends *dist=100* means that edit distance threshold were 100. Since not one metabolite had a name longer than 100 characters, it can be assumed that this threshold is disabled. *Dist=20* means that for each particular pair of metabolites it is allowed to have 20 different characters. *Dist=5* means that only 5 edit operations are allowed to edit one name into another - if the edit distance for particular pair of metabolites is longer, the pair is automatically discarded.

In fact the edit distance criterion removes pairs with high similarity ratio if the number of different symbols exceeds the edit distance. Therefore the edit distance criterion gives effect when the similarity ratio drops below 80% because in case of high similarity usually the number of different letters is low. At some level of similarity ratio the edit distance criterion gives impact and prevents inclusion of metabolite pairs with high number of different letters.

The results revealed by automatic metabolite mapping were manually curated by a biologist. To reduce the amount of data for manual curation, the filtering by compartments were used (because such information was available). During the manual curation the biologist approved 407 metabolites out of 289 automatically mapped. Data used in manual curation is available in supplementary materials.

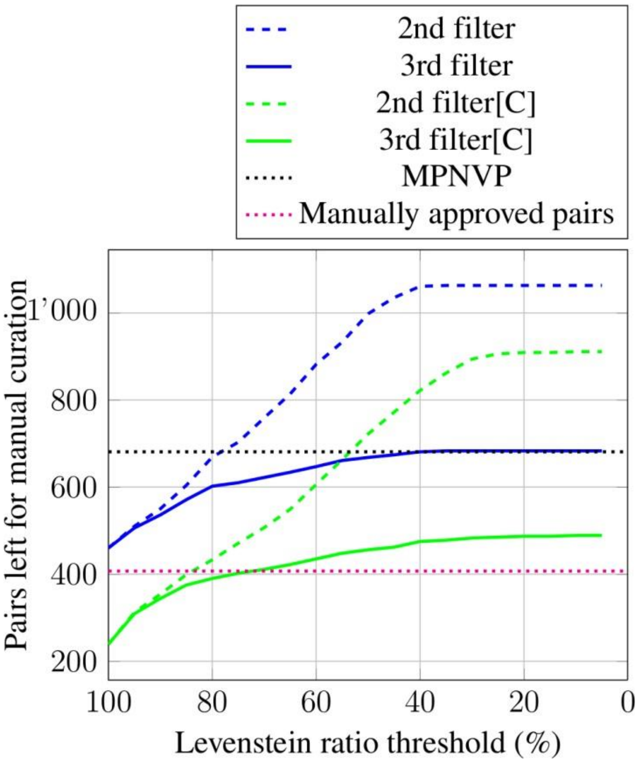

Fig. 6. **The effect of filtering by compartments on 2nd and 3rd filter.**

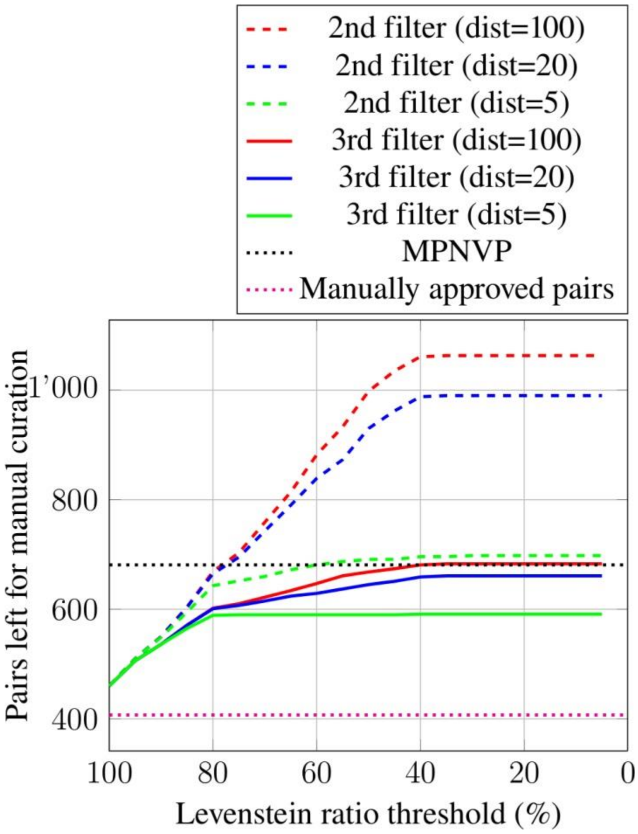

Fig. 7. **The effect of combined thresholds: similarity ratio and edit distance**

4. Conclusion

In case of two model comparison by metabolite names, the total number of combinations to be compared equals to the product of number of metabolites in both models which may lead to thousands or millions of candidate pairs. Therefore manual comparison is not a good alternative. The comparison by name similarity is a compromise between automatic and fully manual comparison. The major part of possible candidate pairs can be rejected automatically using simple, but laborious inspection. While the *three-level-filtering* algorithm does not solve the problem of fully automatic comparison, it can drastically reduce the amount of data that requires manual curation.

Low similarity ratio threshold values (down to 0%) leave all the work for manual comparison while high values (up to 100%) take into account only identical names. Therefore in case of rough comparison (for instance when many models have to be compared) the similarity ratio threshold should be kept high to reduce manual curation workload. In case of detailed comparison the similarity ratio threshold should be kept lower (50-60%) to avoid rejection of potential metabolite pairs.

Compartments should be taken into account mapping metabolites when possible to reduce the manual curation. The introduction of edit distance helps to filter away metabolite pairs with number of different symbols above the threshold. The effect of edit distance increase in case of long metabolite names.

The use of additional criteria (compartments, formulas, edit distance) can only improve the quality of automatic metabolites reconciliation, however, most of this data is often included in models.

Acknowledgements

This work and academic study is funded by a project “Support for doctoral studies in LUA” /2009/0180/1DP/1.1.2.1.2/09/IPIA/VIAA/017 agreement Nr. 04.4-08/EF2.D3.30

References

Boele, J., Olivier, B.G. & Teusink, B. (2012). FAME, the Flux Analysis and Modeling Environment. *BMC systems biology*, 6(1), p.8. http://dx.doi.org/10.1186/1752-0509-6-8

Duarte, N.C., Herrgård, M.J. & Palsson, B.Ø. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7), pp.1298–309. http://dx.doi.org/10.1101/gr.2250904

Gay, S., Soliman, S. & Fages, F. (2010). A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18), pp.i575–i581. http://dx.doi.org/10.1093/bioinformatics/btq388

Ideker, T. et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518), pp.929–34. http://dx.doi.org/10.1126/science.292.5518.929

Krause, F. et al. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics* (Oxford, England), 26(3), pp.421–2. http://dx.doi.org/10.1093/bioinformatics/btp642

Kuepfer, L., Sauer, U. & Blank, L.M. (2005). Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome research*, 15(10), pp.1421–30. http://dx.doi.org/10.1101/gr.3992505

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady, 10(8), pp.707–710.

Mednis, M. & Aurich, M.K. (2012). Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models. *Bit-Journal*, 1(1), pp.14–18. http://dx.doi.org/10.11592/bit.121102

Mednis, M., Brusbardis, V. & Galvanauskas, V. (2012). Comparison of genome-scale reconstructions using ModeRator. In 13th IEEE International Symposium on Computational Intelligence and Informatics. Budapest, pp. 79–84.

Mednis, M., Rove, Z. & Galvanauskas, V. (2012). ModeRator - a software tool for comparison of stoichiometric models. In 7th IEEE International Symposium on Applied Computational Intelligence and Informatics. Timisoara, pp. 97–100.

Mulligan, C. (2013). pylevenshtein. Available at: http://code.google.com/p/pylevenshtein/.

Oberhardt, M.A. et al. (2011). Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis P. E. Bourne, ed. *PLoS Computational Biology*, 7(3), p.18. http://dx.plos.org/10.1371/journal.pcbi.1001116

Palsson, B.Ø. (2006). Systems Biology: Properties of reconstructed networks, *Cambridge University Press*. http://dx.doi.org/10.1017/CBO9780511790515

Poggendorff, J.C. (1830). Annalen der Physik, J.A. Barth. Radulescu, O. et al., 2008. Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1), p.86. http://dx.doi.org/10.1186/1752-0509-2-86

Radulescu, O. et al. (2008). Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1), p.86.

Schellenberger, J. et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: COBRA Toolbox 2 . 0. *Nature Protocols*, 6(9), pp.1290–1307. http://dx.doi.org/10.1038/nprot.2011.308

Schulz, M. et al. (2006). SBMLmerge, a system for combining biochemical network models. Genome informatics International Conference on *Genome Informatics*, 17(1), pp.62–71.

Thiele, I. et al. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, (September 2012).

Thiele, I. & Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1), pp.93–121. http://dx.doi.org/10.1038/nprot.2009.203

Biosystems and Information Technology (2013) Vol.2(1) 6-10
DOI: http://dx.doi.org/10.11592/bit.130502

original scientific article

Information processing for remote recognition of the state of bee colonies and apiaries in precision beekeeping (apiculture)

Aleksejs Zacepins[1,2*], **Egils Stalidzans**[2,3]

[1]*Computer Control Group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV-3001, Jelgava, Latvia*
[2] *SIA TIBIT, Dobeles iela 10-9, LV-3001, Jelgava, Latvia*
[3]*Biosystems Group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV-3001, Jelgava, Latvia*
**Corresponding author*
alzpostbox@gmail.com

Received: 25 May 2013; accepted: 30 May 2013; published online: 30 May 2013.
This paper has no supplementary material.

Abstract: *The principles of Precision Agriculture (PA) can be applied also to the beekeeping branch. Precision Beekeeping (PB) (Precision Apiculture) can be implemented as a three phase cycle including 1) data collection, 2) data analysis and 3) application. The first two phases are based information technologies in case of remote recognition. The third phase is realized manually according to the decisions made after data collection and analysis.*

This study is dedicated to the information processing approaches taking into account the peculiarities of the beekeeping branch. Classification of deviations at several levels is proposed: colony level (most colonies in the same location behave normally); apiary level (most apiaries at other locations behave normally); bee farm level (most apiaries of other bee farms behave normally) and regional level (most bee farms in the region do not behave normally). Two levels of information analysis are suggested: bee farm level (information about individual colonies in different apiaries) and regional level (summary of information collected at the level of bee farms).

Decision support systems (DSS) are proposed to automate data analysis. Continuous operation and high processing capacities of electronics can significantly improve implementations of PB. DSS may be delegated to make some decisions automatically or request the analysis of proposed decision by a specialist in data processing or beekeeper.

Keywords: precision beekeeping, precision apiculture, data collection, decision support system.

1. Introduction

Information and communication technologies (ICT) provide indispensable support for business, agriculture, and production processes. The rapid development of information technologies and computer control enabled the development of precision agriculture (PA) aiming to monitor and control individual agricultural units. The definition of Precision Agriculture is still developing and improving, because technologies that are used in PA are changing and comprehension about theoretical and practical opportunities is developing. Over the years the emphasis of the definition has changed from following soil characteristics in agriculture (Robert and Stafford, 1999) to more complicated where quality of the end product and impact on the environment becomes more relevant (McBratney et al., 2005).

PA principles have been adapted to several agricultural (McBratney et al., 2005; Morais et al., 2008; Whelan and McBratney, 2000) and forestry (Zhang et al., 2011) branches. The same principles can be applied also for beekeeping taking a bee colony as the smallest industrial unit of interest in beekeeping. Apiculture (Beekeeping) is one of the branches of agriculture where precision approach is recently adapted (Zacepins et al., 2012). Precision beekeeping (PB) approach is based on the continuous measurements of individual bee colonies and can be applied all year round thus detecting different states of colonies and apiaries enabling rapid reaction by the beekeeper in case of necessity (Zacepins et al., 2012).

PA branches can be analyzed as a three phase cycle including 1) data collection, 2) data analysis and 3) application (Terry, 2006). These phases are at very different development stage in case of PB. Thre first two phases are closely related to the information technologies while the third one usually has to be done by a beekeeper according to the decisions made after data anlysis.

There are quite many parameters that can be measured to assess the state of individual bee colonies. Still they are very different in terms of information processing and transmission. For instance a temperature measurement returns just one number that can be easily stored in memory or transmitted while sound measurements request intensive processing or transmission of large amount of data.

The data analysis phase is the stumbling block to adoption of PA generally (McBratney et al., 2005). The same applies to the precision beekeeping. Some data analysis based decision support systems are reported in the literature. Most of them concentrate on single colony level while the others are aiming for benefit of larger regions involving wider community into measurements and data exchange enabled by information technology.

The article concentrates on the remote performance of measurements, data analysis of measurements and principles of decision support systems taking into account the peculiarities of beekeeping branch and opportunities offered by information technologies.

2. Different level data based state recognition in precision beekeeping

Compared to other branches of agriculture the industrial beekeeping has several peculiarities which should be addressed by PB technologies. (1) Honey bees are social insect and one industrial unit is a bee colony that consists of tens of thousands of individual bees. (2) The foraging area of honey bees is around their location within radius of about 3 kilometres and beekeeper can influence the feedstock by transporting bee colonies to places with different nectar sources. (3) Bee colonies usually are kept in groups with limited number of 10-30 colonies in one location because more colonies may not have sufficient amount of nectar available in the foraging area which leads to reduced incomes per colony. (4) Remote state recognition is important because bee colonies in apiaries can be left without inspection for long time if they are in acceptable state. (5) Wide foraging area adds complexity to the control of bee diseases.

A beekeeper is interested in classification of deviations at several levels taking into account the above mentioned peculiarities and business tasks of beekeeping (Fig. 1): colony level (most colonies in the same location behave normally); apiary level (most apiaries at other locations behave normally); beekeeper’s farm level (most apiaries of other bee farms behave normally) and regional level (most bee farms in the region do not behave normally).

To operate with data at different levels it is necessary to centralize the data by it’s transfer using internet or other transmission technologies depending on the local circumstances to extract maximal benefit from any measurement. The value of a single measurement may increase analyzing it in context with other ones.

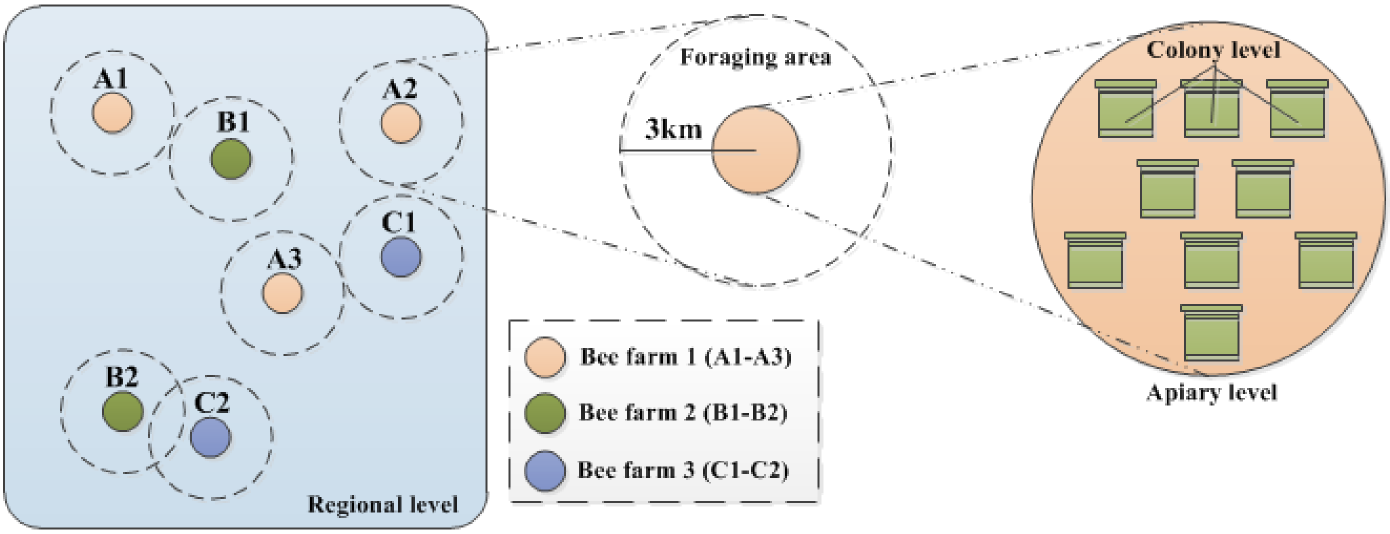

Fig. 1. **Different scales of information collection in beekeeping: colony level, apiary level, bee farm level and regional level.**

2.1. Colony level decisions

Colony level decisions should be made based on individual colony measurements and monitoring. For example, based on low temperature in a colony it is possible to conclude if the bee colony is in a passive/inactive state, if other colonies have temperature about 30°C (Stalidzans and Berzonis, 2013). Preswarming and swarming state detection is another colony-level challenge for automatic remote detection systems.

Temperature measurements of the individual colonies seem to be the most cost efficient way to monitor colony activity and behaviour (Zacepins and Karasha, 2013; Zacepins et al., 2013). Other parameters like air humidity, gas content, sound, video and may be used as well. Still analysis of economical feasibility of different systems has to be clarified depending on technological, climatic and genetic context of particular bee farms or even apiaries.

2.2. Apiary level decisions

Apiary level problems are mainly related to the location of apiary assuming that all the apiaries of particular bee farm are treated in the same way. In this case all the apiary colonies are exposed to the apiary specific factor. Some examples of apiary level factors are: limitations of nectar availability, application of pesticides within the foraging area, noise or other disturbances close to the apiary, theft, diseases.

In spite of the fact that all the colonies in apiary are exposed to the disturbing factor their reaction may be different depending on the internal state of colonies (after swarming, queenless etc.). Video technologies can be used for apiary level monitoring to observe the whole apiary. Different approaches of video activation can be used to reduce the amount of produced data (Meitalovs et al., 2009) if necessary. Climate observation tools with remote connectin can be applied to determine the local weather parameters.

According to the measurements apiary level decisions can be very different: visit of the apiary to examine the situation in details, transportation of bee colonies to a different place, feeding of bee colonies, disease treatment etc.

2.3. Bee farm level decisions

Farm specific problems mostly are caused by the way of operation of bee farm and should be observed in all the apiaries belonging to the same bee farm (Fig. 1). The causes should be of technological origin: wrong timing of operations, inefficient medical treatment etc.

The decisions should be based on analysis of the applied technologies and approaches if similar deviations are not observed at apiaries of a different bee farms having apiaries in the same area which should be exposed to similar circumstances (for instance B2 and C2 in Fig.1). Thus it is critical for farm level decisions to have access to the measurements of other farms in the same region to distinguish between bee farm and regional level problems as they may require different decisions and actions.

One example of remote decision system is practically applied for indoor wintering of bees where temperature monitoring at individual colony level is proposed (Zacepins and Stalidzans, 2012). This kind of architecture can be used both for apiary and farm level (Fig. 2).

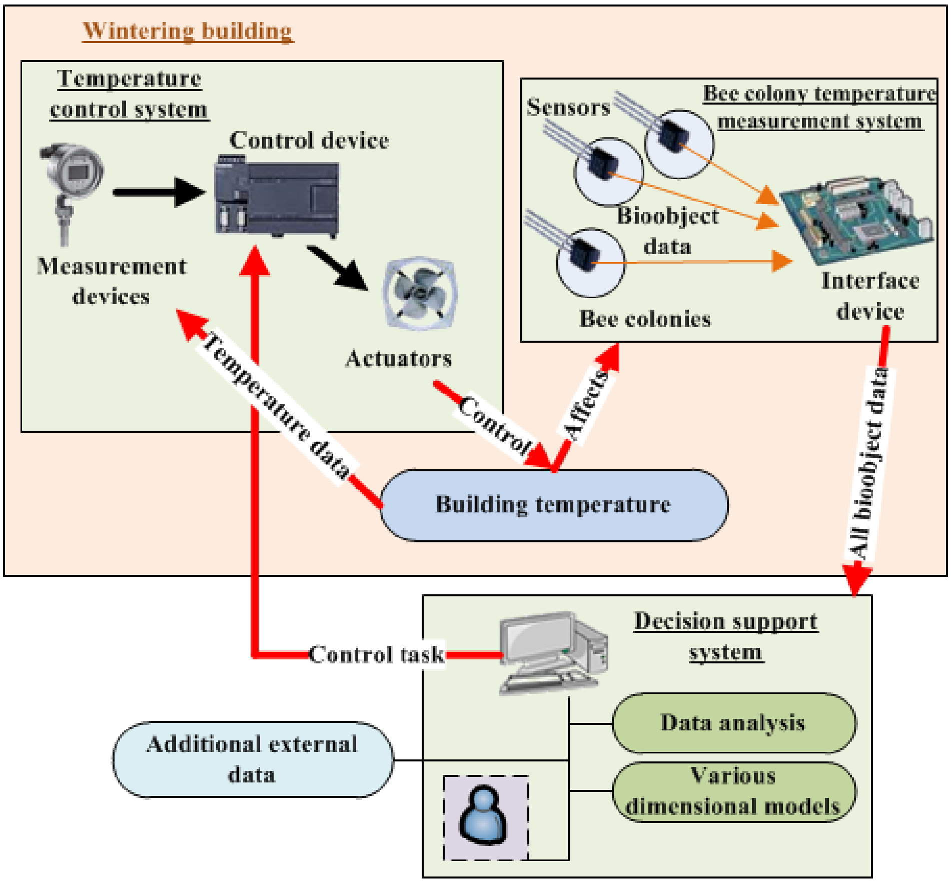

Fig. 2. **Architecture of decision support system for bee wintering building.**

2.4. Regional level based decisions

Similar deviations from normal behaviour of significant part of apiaries of different farms in the same geographical region are a signal for necessary measures at regional scale. Regional level problems can be caused by unusual climatic circumstances (dry, wet, cold, hot), diseases of bees or plants etc. Information about regional problems, detected by collaboration of several bee farms by the exchange of measurements, can be spread among all the beekeepers (including hobby ones) in the region even if they do not participate in the collection of measurements. Informing about regional problems some activities can be suggested to minimize the impact of discovered regional problems.

Early diagnostics of diseases, especially infectious ones can prevent significant losses in a region directly for beekeepers and indirectly for agriculturists due to reduced pollination. Colony collapse disorder (CCD) may serve as a good example for regional level problems (Cox-Foster et al., 2007; Van Engelsdorp et al., 2008).

Example of functioning regional level data collection system prototype is transnational bee colony monitoring system which operates as an international network including bee colonies in Denmark, Sweden, Norway, Latvia and Germany (http://biavl.volatus.de/bsm0/BSM.html#). The advantage of this system is it's geographical coverage and the number of measured parameters: weight changes, ambient temperature, precipitation, and the temperature of the colony. A disadvantage is the low number of monitored colonies and lack of decision support system. Independent on it's current execution and intensity of use this system is a good prototype for future developments.

Another regional level project is NASA Goddard Space Flight Center initiated project (http://honeybeenet.gsfc.nasa.gov/) where daily weighing of hives by volunteers is merged with satellite data (Nightingale et al., 2008). Beekeepers can also directly monitor the weight changes to estimate the amount of incoming nectar.

Mentioned projects indicate both interest in remote data collection and sharing and technical opportunities for practical implementation at a regional level. Still reliable decision support systems are needed to make use of collected data.

3. ICT aspects of information collection and decision support in PB

3.1. Information collection, processing and transfer

A prerequisite of the above mentioned four level based decisions is a system of information collection, processing and transfer. Two levels of information collection can be used: bee farm level (information about individual colonies in different apiaries) and regional level (summary of information at the level of bee farms). Bee farm level information collection is reasonable as the lowest level of decision making assuming that all the apiaries of particular bee farm is managed by the same team. The collection and analysis of regional information can be performed by regional governments, beekeeper societies or temporary projects.

While developing farm level systems it is crucial to decide about information processing and transfer options, because it is possible to process data onsite and transfer just a summary or transfer all the raw data to a remote computational centre for further processing. The decision about local or centralized information processing can significantly influence the costs of system. That can be very important issue depending on the processing peculiarities of particular parameter. For instance, the result of temperature or humidity measurements is just a digit that does not request much processing and even transfer of that information is cheap. That is different in case of sound measurements or video recording where both processing and transfer are much more complicated and costly.

Apiaries which are located in sites without centralized electricity supply have another problem: energy source. Depending on a solution (batteries, solar panel, wind generator etc.) additional limitations may appear having impact on the feasibility of PB system of interest (Zacepins et al., 2013).

3.2. Decision support systems (DSS)

The second stage of PA approach – data analysis – can be performed by the beekeeper interpreting received data. Information technologies can at least partly replace a specialist in case of large data amount or if continuous analysis is necessary. such approach could help beekeepers which hardly could interpret the data by themselves. The computational support of beekeepers can be done using DSS where different algorithms and models can be implemented (Fig. 3).

Task of the models is to represent various real physical, biological, economical or other processes. Usually models give a simplified view about process, but nevertheless information, which is provided by the model, is useful for the detailed research of the process (Sokolowski and Banks, 2009). Models can be divided in two categories: identification (qualitative) and quantitative models (Holjushkin and Grazhdannikov, 2000). For instance, identification model could be used to determine if colony is in the preswarming state where the answer is “yes” or “no”. A quantitative model would predict, for instance, the number of bees in the colony in particular date where the answer is a number of bees.

DSS may use different combinations of different model types to suggest particular decisions to the beekeeper.

Authors propose to divide decision making process in three levels (Fig. 4):

- *input data level* – where all needed data about process and object should be defined;
- *model level* – where input data is used by various different dimension models with main aim to determine the object state and status of the process;
- *decision level* – where model outputs are analysed with main aim to choose the right decision (beekeeping operation to be performed).

Different states of colony, apiary, bee farm and region can be recognized with different level of reliability. Therefore, depending on the importance of detected state and importance of immediate action DSS may be delegated to make some decisions automatically or request the analysis of proposed decision by a specialist in data processing or beekeeper. Thus the combination of ICT applications in data collection and data analysis can give important new tools for PB applications at bee farm and regional level.

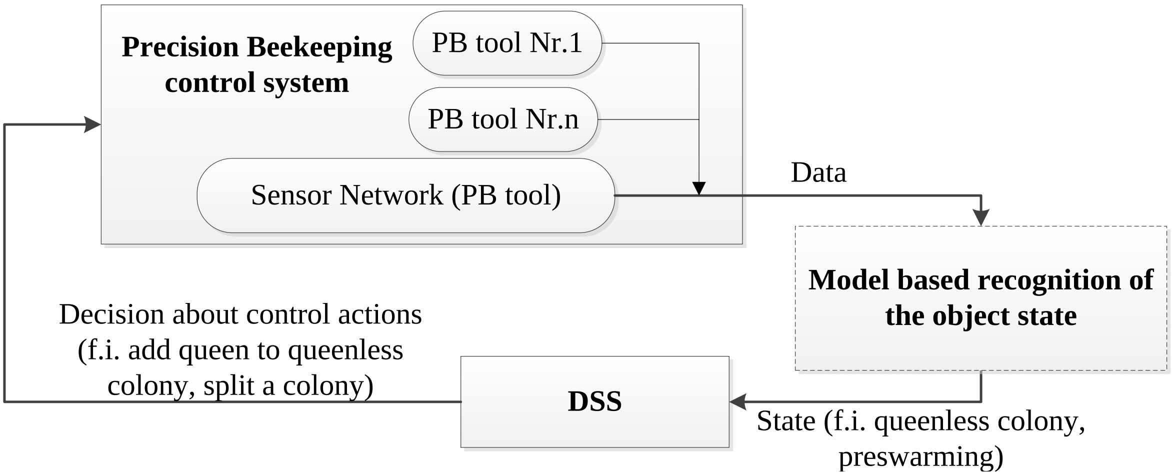
Precision Beekeeping control system
PB tool Nr.1
PB tool Nr.n
Sensor Network (PB tool)
Data
Model based recognition of the object state
DSS
State (f.i. queenless colony, preswarming)
Decision about control actions (f.i. add queen to queenless colony, split a colony)

Fig. 3. **Diagram of information flow for DSS implementation in PB.**

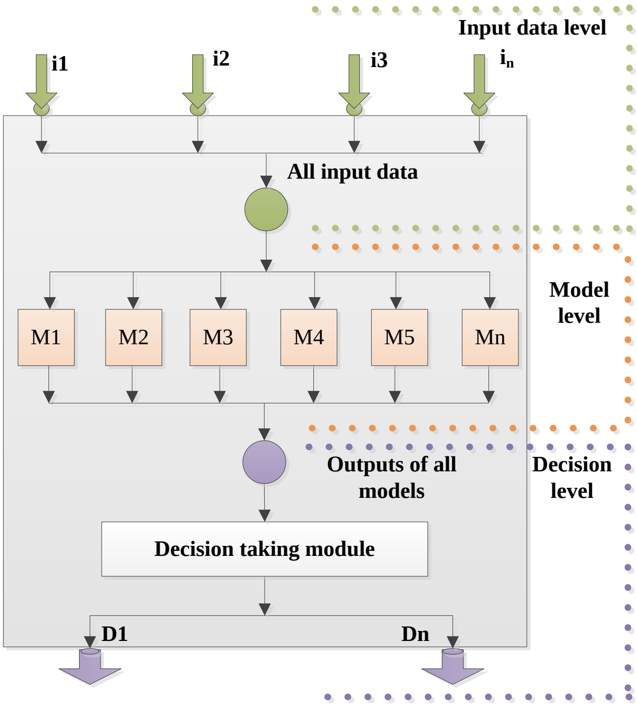

Fig. 4. **Three levels of the decision making in DSS.**

4. Conclusion

ICT can be applied at two phases of Precision Beekeeping (PB): data collection and data analysis. Data collection includes the data collection on the colony level, data processing and data transfer to make them available for the specialist or decision support system. Rational compromise between local processing of information and transmission of unprocessed information has to be found depending on several factors.

Decision support systems (DSS) are proposed to automate data analysis. Continuous operation and high processing capacities of electronics thus can be useful help in PB. DSS may be delegated to make some decisions automatically or request the analysis of proposed decision by a specialist in data processing or beekeeper.

Several levels of decisions can be made using PB approach based on different level of information: colony level, apiary level, bee farm level and regional level. Two levels of information analysis can be used: bee farm level (information about individual colonies in different apiaries) and regional level (summary of information collected at the level of bee farms). Bee farm level information collection is reasonable as the lowest level of decision assuming that all the apiaries of particular bee farm is managed by the same team. The collection and analysis of regional information can be performed by regional governments, beekeeper societies or temporary projects.

Three level decision making process is proposed: input data level, model level and decision level. It is proposed to use interaction between quantitative and qualitative models.

References

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., Moran, N. a, Quan, P.-L., et al. (2007), “A metagenomic survey of microbes in honey bee colony collapse disorder.,” *Science (New York, N.Y.)*, Vol. 318 No. 5848, pp. 283–7. http://dx.doi.org/10.1126/science.1146498

Van Engelsdorp, D., Hayes, J., Underwood, R.M. and Pettis, J. (2008), “A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008.,” *PloS one*, Vol. 3 No. 12, p. e4071. http://dx.doi.org/10.1371/journal.pone.0004071

Holjushkin, J.P. and Grazhdannikov, E.D. (2000), “Sistemnaja klassifikacija arheologicheskoj nauki,” *Klassifikacionnue modeli v arheologicheskom naukovedenie*.

McBratney, A., Whelan, B., Ancev, T. and Bouma, J. (2005), “Future directions of precision agriculture,” *Precision Agriculture*, Vol. 6, pp. 7–23.

Meitalovs, J., Histjajevs, A. and Stalidzans, E. (2009), “Automatic Microclimate Controlled Beehive Observation System,” *8th International Scientific Confernce “Engitieering for Rural Development”*, Jelgava, Latvia, Latvia University of Agriculture, pp. 265–271.

Morais, R., Fernandes, M. a., Matos, S.G., Serôdio, C., Ferreira, P.J.S.G. and Reis, M.J.C.S. (2008), “A ZigBee multi-powered wireless acquisition device for remote sensing applications in precision viticulture,” *Computers and Electronics in Agriculture*, Vol. 62 No. 2, pp. 94–106. http://dx.doi.org/10.1016/j.compag.2007.12.004

Nightingale, J.M., Esaias, W.E., Wolfe, R.E., Nickeson, J.E. and Ma, P.L.A. (2008), “Assessing Honey Bee Equilibrium Range and Forage Supply using Satelite-Derived Phenology,” *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, pp. III – 763–III – 766. http://dx.doi.org/10.1109/IGARSS.2008.4779460

Robert, P. and Stafford, J. (1999), “Precision Agriculture: research needs and status in the USA.,” *Conference on Precision Agriculture*, pp. 19–33.

Sokolowski, J.A. and Banks, C.M. (2009), *Principles of Modeling and Simulation*, New York, Wiley, p. 257. http://dx.doi.org/10.1002/9780470403563.index

Stalidzans, E. and Berzonis, A. (2013), “Temperature changes above the upper hive body reveal the annual development periods of honey bee colonies,” *Computers and Electronics in Agriculture*, Vol. 90, pp. 1–6. http://dx.doi.org/10.1016/j.compag.2012.10.003

Terry, B. (2006), *Precision Agriculture*, Thomson Delmar learning, p. 224.

Whelan, B.M and McBratney, A.B. (2000), “The ‘null hypothesis’ of precision agriculture management,” *Precision Agriculture*, Vol. 2, pp. 265–279.

Zacepins, A. and Karasha, T. (2013), “Application of temperature measurements for the bee colony monitoring : a review,” *Proceedings of the 12th International Scientific Conference “Engineering for Rural Development”*, Jelgava, Latvia, pp. 126–131.

Zacepins, A., Stalidzans, E. and Karasha, T. (2013), “Profitability ranking of precision agriculture measurement systems implementation,” *Proceedings of the 12th International Scientific Conference “Engineering for Rural Development”*, Jelgava, Latvia, pp. 164–169.

Zacepins, A. and Stalidzans, E. (2012), “Architecture of automatized control system for honey bee indoor wintering process monitoring and control,” *Proceedings of 13th International Carpathian Control Conference (ICCC), 28-31 May 2012*, High Tatras, pp. 772–775.

Zacepins, A., Stalidzans, E. and Meitalovs, J. (2012), “Application of Information Technologies in Precision Apiculture,” *Proceedings of the 11th International Conference on Precision Agriculture, 15-18 July 2012*, Indianapolis, IN, USA.

Zhang, Q., Li, J. and Rong, J. (2011), “Application of WSN in precision forestry,” *IEEE 2011 10th International Conference on Electronic Measurement & Instruments*, IEEE, pp. 320–323. http://dx.doi.org/10.1109/ICEMI.2011.6038006

Biosystems and Information Technology (2013) Vol.2(1) 11-14
DOI: http://dx.doi.org/10.11592/bit.130503

original scientific article

Determinition of best set of adjustable parameters with full search and limited search methods

Ivars Mozga

SIA TIBIT, Dobeles iela 10-9, LV-3001, Jelgava, Latvia
ivars.m@gmail.com

Received: 26 May 2013; accepted: 30 May 2013; published online: 31 May 2013.
This paper has no supplementary material.

Abstract: *In case of optimization of biochemical networks the best combination of adjustable parameters has to be found keeping low the costs of modification of biochemical network and reducing the risk of unpredicted side effects of processes outside the scope of the model. Dynamic models of biochemical networks usually are in form of system of differential equations and therefore can not be optimized analytically. Usually they are optimized using computationally demanding global stochastic optimization methods which may have hardly predictable duration of optimization. Optimization of all the possible combinations is necessary to find the best set of adjustable parameters per number of parameters in combination. Due to the combinatorial explosion the full search often is not feasible approach if computational and time resources are limited.*

In this study the performance of full search is compared with forward selection and three metabolic control based methods which need significantly less combinations to be examined. It is found that forward selection is a good compromise in case of larger number of adjustable parameters where the number of possible combinations exceeds the available resources needed for full search in the adjustable parameter space. It is proposed to combine full search for small number of adjustable parameters with forward selection at larger number of adjustable parameters per combination.

Keywords: ranking, adjustable parameters, optimization, dynamic model, biochemical networks.

1. Introduction

The development of models of biochemical processes becomes more common task in the developing fields of systems biology and synthetic biology which in turn help in the development of biotechnological, medical and environmental solutions. In case of dynamic models parameter estimation (Mendes et al., 2009; Moles et al., 2003; Rodriguez-Fernandez et al., 2006) is often used to find the correct values of model parameters to repeat the process of interest. Design task (Mendes and Kell, 1998) is performed to find the most effective manipulation with adjustable parameters of the process of interest. Global stochastic optimization methods are often used to solve both kind of tasks (Banga, 2008; Rodriguez-Fernandez et al., 2006). Some disadvantages of global stochastic optimization methods are 1) the hardly predictable duration of optimization (Mozga and Stalidzans, 2011a; Nikolaev, 2010) and 2) possible stagnation in local optima (Mozga and Stalidzans, 2011b; Sulins and Mednis, 2012).

Often both in parameter estimation and design tasks it is necessary to find the smallest set of adjustable parameters to satisfy the task setting: requested level of similarity between experimental measurements and the behaviour of model in case of parameter estimation task or the necessary level of improvement of objective function in case of design task. The problem appears because of the combinatorial nature of the task when the best combination of limited number of parameters has to be found (Pharkya and Maranas, 2006; Yousofshahi et al., 2013). One of approaches is to use the optimization potential to find out when further improvement becomes impossible (Mozga and Stalidzans, 2011c; Mozga, 2012). Application of parallel optimization runs is proposed to analyze (Kostromins et al., 2012) and automatically detect the moment of termination of optimization runs (Sulins and Stalidzans, 2012). Still those attempts remain computationally consuming and request analysis of large amount of combinations.

There are several approaches based on some metabolic network specific features. A method that uses all the enzymes on the way between substrate and product is proposed by Kacser and Acerenza (Kacser and Acerenza, 1993). The advantage of this approach is it’s simplicity. The main disadvantage is the fact that good increase of objective function can be reached also using smaller set of enzymes (Nikolaev, 2010; Rodríguez-Prados et al., 2009). Metabolic Control Analysis (MCA) approach (Fell, 1992) based solutions are proposed (Acerenza and Ortega, 2007; Hatzimanikatis, 1999; Magnus et al., 2009; Rodríguez-Prados et al., 2009; Stephanopoulos and Simpson, 1997). Still most of them need transformations of models that is hard to automate.

In this study several simple approaches including MCA based ones are tested on their efficiency and compared with full combinatorial search. The tested approaches examine small number of combinations which is close to the number of adjustable parameters. The execution of all tested methods can be performed without transformations of models and can be automated. Tests are performed using model of yeast glycolysis (Hynne et al., 2001).

2. Materials and methods

2.1. Optimization task and software

Yeast glycolysis model (Hynne et al., 2001) downloaded from Biomodels data base (Le Novère et al., 2006) is used as a test model for optimization. The model contains 2 compartments, 24 reactions and 25 metabolites. Objective function in all optimization runs was

$$K = \frac{Ethanol\ flow}{Glucose uptake} + 5 * Ethanol\ flow$$

Concentrations of enzymes catalyzing 15 reactions were chosen as total set of adjustable parameters. COPASI (Hoops et al., 2006), build 30, is used as optimization tool. Particle swarm optimization method is applied with following method parameters: Iteration Limit: 2000; Swarm Size: 50; Std. Deviation: 1e-06; Random Number Generator: 1; Seed: 0. The values of adjustable parameters were allowed to change within a wide range from -99% up to 1000% from their initial values. "Steady state" subtask of optimization was chosen to reject solutions without steady state. ConvAn software (Kostromins et al., 2012) is used for analysis of optimization dynamics.

2.2. Ranking methods

Computationally demanding full combinatorial search of the possible space of adjustable parameter combinations is compared to several alternatives: forward selection method and three MCA based methods. The proposed MCA methods are applicable without any alterations or analysis of the biochemical network like other MCA based methods (Acerenza and Ortega, 2007; Hatzimanikatis, 1999; Magnus et al., 2009; Stephanopoulos and Simpson, 1997) bringing the advantage of possible automation of the model analysis as a single task. Flux control coefficients (FCC) or concentration control coefficients (CCC) are used depending on the expression of objective function. Full combinatorial search in this case serves as a benchmark for evaluation of other methods.

2.2.1. 1-st method

The 1st method starts with no adjustable parameters in the model and adds the next adjustable parameter with the largest module of flux control coefficient (FCC) of initial model (before optimizations) that corresponds to the objective function. That is repeated until all the adjustable parameters are involved. In each step the values of adjustable parameters reached in the previous step remain fixed. The number of tested combinations in this case corresponds to the number of adjustable parameters. In case of proposed model that is 15.

2.2.2. 2-nd method

The 2nd method is similar to the 1st method. The only difference is that the FCC coefficients of a steady state after optimization of all adjustable parameters in one set are taken into account calculating the largest module of FCC that corresponds to the objective function. This method takes into account the best possible improvement of the model instead of the initial state of the model which is not changed under influence of the objective function. The number of tested combinations in this case corresponds to the number of adjustable parameters. In case of proposed model that is 15.

2.2.3. 3-rd method

The 3rd method is reverse compared to the 2nd one: the starting point is a model which is optimized with all the adjustable parameters in one set. FCC of the optimized model is used. The main difference is that the initial state is when all the adjustable parameters are included and in each step the adjustable parameter with the smallest module of FCC that corresponds to the objective function is removed. That is similar to the backward elimination approach. The process is completed when all the parameters are removed from the set of adjustable parameters. The number of tested combinations in this case also equals to the number of adjustable parameters (15).

2.2.4. Forward selection

Forward selection method starts with no adjustable parameters in the model and tests the addition of each adjustable parameter using objective function as criterion adding the adjustable parameter that improves the objective function the most, and repeating this process until all adjustable parameters are involved. The number of optimized combinations for n adjustable parameters is equal to their sum:

$$\sum_{i=1}^{n} x_i$$

That is significantly bigger number than the one for 1st, 2nd and 3rd methods. In the experiment only up to eight adjustable parameters per combination were examined because the maximal possible value of objective function (equals to the one of all 15 parameters) was already reached and further addition of adjustable parameters could not increase it.

2.2.5. Full search

Fulle search is the most demanding approach in terms of computational costs because of the large number of combinations. The number of combinations of m parameters out of n parameters can be calculated by formula:

$$C_n^m = \frac{n!}{m!(n-m)!}$$

The number of all possible combinations with up to n adjustable parameters out of n can be calculated as follows:

$$L_n = \sum_{m=1}^{n} C_n^m = \sum\nolimits_{m=1}^{n} \frac{n!}{m!\,(n-m)!}$$

Therefore the number of combinations of parameters is increasing very fast and reaches 32767 possible combinations for up to 15 parameters out of the set consisting of 15 parameters.

Thus the screening of all the combinations with optimization runs is very demanding in terms of time and computational resources. complicated or almost impossible due to the huge number of combinations and unpredictable optimization time to reach the global optimum close objective function value (Mozga and Stalidzans, 2011a). The main advantage is that checking all the possible combinations it is guaranteed that the best one is found.

Due to the high number of combinations this method is performed for only up to three adjustable parameters. That means 15 combinations of one parameter, 105 combinations of two parameters and 455 combinations of three parameters (totally 475 combinations) are examined.

3. Results and discussion

The best values of objective function per number of adjustable parameters in combinations (Fig. 1) demonstrate big advantage of full search method over the other ones at small number of adjustable parameters. The main drawback is the very high number of searched combinations (475) compared to 3, 3, 114 and 6 combinations for 1st, 2nd, 3rd and forward selection methods.

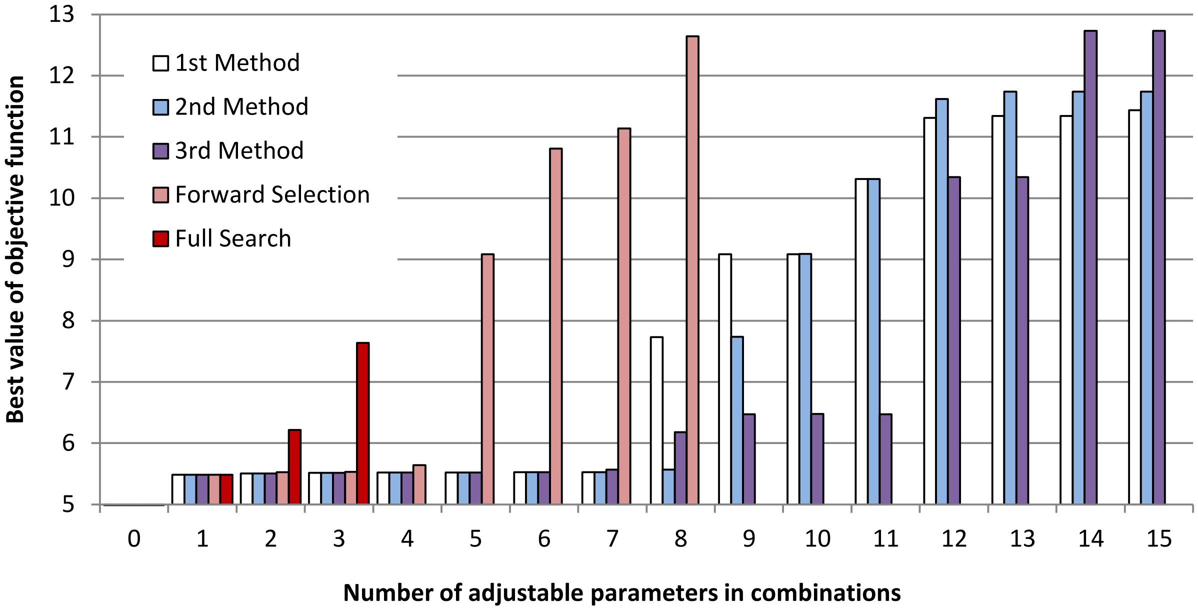

Fig. 1. **The best values of objective function per number of adjustable parameters in combinations for compared methods. Forward selection method is applied for up to eight parameters. Full search is applied for up to three parameters.**

Forward selection method needs 120 combinations for 15 adjustable parameters and show convincing advantage over all the MCA based methods which request almost 10 times less (15) combinations to cover all the 15 parameters. At the same time the maximal possible increase of objective function is reached already including eight parameters and correspondingly using 92 instead of 120 combinations. Thus the experiment shows clear correlation between the performance of method and the number of examined combinations.

Comparing the MCA based methods the 1st and 2nd method show very similar behavior except for 8 and 9 parameters. The 3rd method has slightly better performance number of adjustable parameters close to the maximum.

Full search for small number of adjustable parameters could be combined with forward selection taking into account the good performance of full search at small number of adjustable parameters and forward selection as the second best method.

4. Conclusion

The tested MCA based or forward selection based methods can not compete with the full search method. Therefore full search method should be used if the computational resources and time are available and high confidence about closeness of the best result to the global optima is needed.

Forward selection is a good compromise in case of larger number of adjustable parameters where the number of possible combinations exceeds the available resources needed for full search in the parameter space.

All MCA based methods demonstrate poor performance and might be outperformed even by selection of random combinations (not tested in this study).

The performed experiments can not be generalized for other optimization tasks and models but it clearly shows that poor performance of tested methods with limited search can have very poor performance. Generally the experiment shows clear correlation between the performance of method and the number of examined combinations.

Full search for small number of adjustable parameters could be combined with forward selection taking into account the good performance of full search at small number of adjustable parameters and forward selection as the second best method.

References

Acerenza, L. and Ortega, F. (2007), “Modular metabolic control analysis of large responses.,” *The FEBS journal*, Vol. 274 No. 1, pp. 188–201. http://dx.doi.org/10.1111/j.1742-4658.2006.05575.x

Banga, J.R. (2008), “Optimization in computational systems biology.,” *BMC systems biology*, Vol. 2, p. 47. http://dx.doi.org/10.1186/1752-0509-2-47

Fell, D.A. (1992), “Metabolic control analysis: a survey of its theoretical and experimental development.,” *Biochemical Journal*, Portland Press Ltd, Vol. 286 No. Pt 2, pp. 313–330. Retrievedfrom http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1132899/

Hatzimanikatis, V. (1999), “Nonlinear metabolic control analysis.,” *Metabolic engineering*, Vol. 1 No. 1, pp. 75–87. http://dx.doi.org/10.1006/mben.1998.0108

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., et al. (2006), “COPASI--a COmplex PAthway SImulator.,” *Bioinformatics (Oxford, England)*, Vol. 22 No. 24, pp. 3067–74. http://dx.doi.org/10.1093/bioinformatics/btl485

Hynne, F., Danø, S. and Sørensen, P.G. (2001), “Full-scale model of glycolysis in Saccharomyces cerevisiae,” *Biophysical Chemistry*, Vol. 94 No. 1-2, pp. 121–163. http://dx.doi.org/10.1016/S0301-4622(01)00229-0

Kacser, H. and Acerenza, L. (1993), “A universal method for achieving increases in metabolite production.,” *European journal of biochemistry / FEBS*, Vol. 216 No. 2, pp. 361–7. http://dx.doi.org/10.1111/j.1432-1033.1993.tb18153.x

Kostromins, A., Mozga, I. and Stalidzans, E. (2012), “ConvAn: a convergence analyzing tool for optimization of biochemical networks,” *Biosystems*, Vol. 108 No. 1-3, pp. 73–77. http://dx.doi.org/10.1016/j.biosystems.2011.12.004

Magnus, J.B., Oldiges, M. and Takors, R. (2009), “The identification of enzyme targets for the optimization of a valine producing Corynebacterium glutamicum strain using a kinetic model,” *Biotechnology progress*, Wiley Online Library, Vol. 25 No. 3, pp. 754–762. doi:10.1021/bp.184

Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J.O. and Kummer, U. (2009), “Computational Modeling of Biochemical Networks Using COPASI,” inMaly,I. V (Ed.),*Systems Biology*, Totowa, NJ, Humana Press, Vol. 500, pp. 17–59. http://dx.doi.org/10.1007/978-1-59745-525-1

Mendes, P. and Kell, D. (1998), “Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation,” *Bioinformatics*, Vol. 14 No. 10, pp. 869–883. http://dx.doi.org/10.1093/bioinformatics/14.10.869

Moles, C.G., Mendes, P. and Banga, J.R. (2003), “Parameter estimation in biochemical pathways: a comparison of global optimization methods.,” (Skjoldebremd,C. and Trystrom,G.,Eds.)*Genome Research*, Goetheborg, Sweden, Vol. 13 No. 11, pp. 2467–2474. http://dx.doi.org/10.1101/gr.1262503

Mozga, I. (2012), *Steady state optimization procedure of biochemical networks*, Latvia University of Agriculture. Retrieved from http://llufb.llu.lv/dissertation-summary/information-technologies/Promocijas_darba_kopsavilkums_IvarsMozga_LLU_ITF_2012.pdf

Mozga, I. and Stalidzans, E. (2011a), “Convergence dynamics of biochemical pathway steady state stochastic global optimization,” *Proceedings of IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), 21-22 November 2011*, Budapest, IEEE, pp. 231–235. http://dx.doi.org/10.1109/CINTI.2011.6108504

Mozga, I. and Stalidzans, E. (2011b), “Convergence Dynamics of Biochemical Models to the Global Optimum,” *Proceedings of the 3rd International Conference on E-Health and Bioengineering, 24-26 November 2011*, Iasi, pp. 227–230.

Mozga, I. and Stalidzans, E. (2011c), “Optimization protocol of biochemical networks for effective collaboration between industry representatives, biologists and modellers,” *Proceedings of nternational Industrial Simulation Conference, 6.-8 June 2011*, Venice, EUROSIS, pp. 91–96.

Nikolaev, E.V. (2010), “The elucidation of metabolic pathways and their improvements using stable optimization of large-scale kinetic models of cellular systems.,” *Metabolic engineering*, Elsevier, Vol. 12 No. 1, pp. 26–38. http://dx.doi.org/10.1016/j.ymben.2009.08.010

Le Novère, N., Bornstein, B.J., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., et al. (2006), “BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.,” *Nucleic acids research*, Vol. 34 No. Database issue, pp. D689–91. http://dx.doi.org/10.1093/nar/gkj092

Pharkya, P. and Maranas, C.D. (2006), “An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems.,” *Metabolic engineering*, Vol. 8 No. 1, pp. 1–13. http://dx.doi.org/10.1016/j.ymben.2005.08.003

Rodriguez-Fernandez, M., Mendes, P. and Banga, J.R. (2006), “A hybrid approach for efficient and robust parameter estimation in biochemical pathways.,” *Bio Systems*, Vol. 83 No. 2-3, pp. 248–65. http://dx.doi.org/10.1016/j.biosystems.2005.06.016

Rodríguez-Prados, J.C., De Atauri, P., Maury, J., Ortega, F., Portais, J.C., Chassagnole, C., Acerenza, L., et al. (2009), “In silico strategy to rationally engineer metabolite production: A case study for threonine in Escherichia coli.,” *Biotechnology and bioengineering*, Vol. 103 No. 3, pp. 609–20. http://dx.doi.org/10.1002/bit.22271

Stephanopoulos, G. and Simpson, T.W. (1997), “Flux amplification in complex metabolic networks,” *Chemical Engineering Science*, Vol. 52 No. 15, pp. 2607–2627. http://dx.doi.org/10.1016/S0009-2509(97)00077-8

Sulins, J. and Mednis, M. (2012), “Automatic termination of parallel optimization runs of stochastic global optimization methods in consensus or stagnation cases,” *Biosystems and Information technology*, Vol. 1 No. 1, pp. 1–5. http://dx.doi.org/10.11592/bit.120501

Sulins, J. and Stalidzans, E. (2012), “Corunner: multiple optimization run manager for Copasi software,” *Proceedings of 5th International Scientific Conference on Applied Information and Communication Technologies, 26-27 April 2012*, Jelgava, pp. 312–316.

Yousofshahi, M., Orshansky, M., Lee, K. and Hassoun, S. (2013), “Probabilistic strain optimization under constraint uncertainty,” *BMC Systems Biology*, Vol. 7 No. 1, p. 29. http://dx.doi.org/10.1186/1752-0509-7-29

Biosystems and Information Technology (2013) Vol.2(1) 15-18
DOI: http://dx.doi.org/10.11592/bit.130504

original scientific article

Automatic termination of parallel optimization runs of global stochastic optimization methods by upper limit criterion

Natalja Bulipopa*, Jurijs Sulins

Biosystems Group, Department of Computer Systems, Latvia University of Agriculture, Liela iela 2, LV-3001, Jelgava, Latvia
**Corresponding author*
cvetkova.natalja@gmail.com

Received: 28 May 2013; accepted: 31 May 2013; published online: 31 May 2013.
This paper has no supplementary material.

Abstract: *Many combinations of adjustable parameters should be tested in optimization experiments of biochemical networks to find the smallest subset of parameters enabling the best improvements of objective function both in case of design task and parameter estimation task. In case of optimization with global stochastic optimization methods one of the problems is the termination of the optimization run looking for a good compromise between spent computational resources and probability that the best found value of objective function will be the global optimum. Longer runs increase the possibility to each the global optimum. Automatic termination criteria in case of consensus or stagnation of parallel optimization runs have been proposed as criteria for automatic termination. Varying the consensus and delay time settings different probability of reaching global optimum and duration of optimization can be reached. It is proposed to modify automatic optimization termination criteria of parallel optimization runs applying upper limit agreement of a number of parallel optimization runs. Automatic application of upper limit agreement would reduce the duration of scanning of the whole space of combination of adjustable parameters. This approach is tested on the yeast glycolysis model with six adjustable parameters using COPASI, CoRunner and ConvAn software for five parallel optimization runs per combination of adjustable parameters.*

Keywords: optimization, termination criteria, global stochastic optimization method, parallel.

1. Introduction

Modelling becomes more and more important part of engineering cycle of biochemical processes (Banga, 2008; Hübner et al., 2011; Mauch et al., 2001). Optimization is one of the application fields of modelling. Improving the performance of a biochemical network for industrial purposes the goal is to use as few as possible alterations to the system to reach industrially interesting strain (Nikolaev, 2010; Pentjuss et al., 2013; Rodríguez-Acosta et al., 1999; Trinh and Srienc, 2009; Unrean et al., 2010).

The necessity to find the best combination per number of adjustable parameters leads to combinatorial explosion of combinations of adjustable parameters which have to be optimized (Stalidzans et al., 2012). Therefore, automatic screening of all the possible combinations becomes important. Unfortunately the systems of differential equations describing dynamics of biochemical networks can not be solved analytically and global computationally expensive stochastic optimization methods are used due to variety of reasons (Banga, 2008; Mendes and Kell, 1998). One of the problems of global stochastic optimization methods is the decision about the termination of optimization run because this kind of methods can not guarantee global optimality (Banga, 2008). That can be compensated by longer optimization runs. The optimization can be terminated when there are no changes of the best value of the objective function for a longer time. Due to stochastic nature the duration of optimization procedure becomes hardly predictable (Mozga et al., 2011; Nikolaev, 2010) even for the same model and constant number of adjustable parameters in combination (Mozga and Stalidzans, 2011a). Application of parallel optimization runs (Sulins and Stalidzans, 2012) with consensus and stagnation criteria is proposed to automate the termination of optimization runs (Sulins and Mednis, 2012).

It is proposed to modify automatic optimization termination criteria (Sulins and Mednis, 2012) of parallel optimization runs (Sulins and Stalidzans, 2012) and use upper limit agreement of a number of parallel optimization runs to reduce the duration of scanning of the whole space of combination of adjustable parameters. This approach is tested on the yeast glycolysis model of Galazzo and Bailey (Galazzo and Bailey, 1990) with six adjustable parameters.

2. Materials and methods

2.1. Model and optimization task setting

The yeast glycolysis model (Galazzo and Bailey, 1990) is used as an optimization task example. The optimization task is set according to the *in silico* optimization experiments of ethanol production performed by Rodriguez-Acosta on the same model (Rodríguez-Acosta et al., 1999). Concentrations of six enzymes catalyzing reactions ATPase, GAPD, Glucose in (Glu), Hexokinase (HK), Phosphofructokinase (PFK) and Pyruvate kinase (PK) are chosen as adjustable parameters. 63 combinations of six adjustable parameters (up to six out of six) are optimized. The range of changes of adjustable parameters is set within range from -99% to +900% (from 100-fold decrease to 10-fold increase) from their initial values. Maximization of the flow through reaction Pyruvate kinase (PK) (proportional to the ethanol production) is set as the objective function. Generally the proposed approach upper limit agreement can be used also for minimization tasks.

2.2. Software tools and optimization settings

The experiments are performed using *COPASI* (Hoops et al., 2006) Build 35 as the optimization tool. Particle swarm optimization method was used with following settings: iteration limit 30000, Swarm Size 50, Std. Deviation 1-e-06, Random Number Generator 1 and Seed 0. Steady state subtask is selected.

CoRunner software (Sulins and Stalidzans, 2012) is used for management of parallel optimization runs of COPASI files. Parallel optimization runs are stopped when all the five parallel runs have reached consensus with consensus corridor 1 % and delay time 15 minutes. The data about the dynamics of the objective function values of these optimization runs are used as test case to examine the reliability of proposed optimization termination criteria which are less demanding than consensus.

ConvAn software (Kostromins et al., 2012) is used for discretization and analysis of convergence dynamics of parallel optimization runs. The discretization step is set at 60 seconds.

2.3. Determination of upper limit

Consensus of two and more optimization runs at the best value of objective function is tested as termination criterion of parallel optimization runs of global stochastic optimization methods (Fig. 1). The best value of objective function for each combination of adjustable parameters after reaching the consensus (as described in section 2.2) is set as global 100% value to assess the effect of less demanding consensus criterion on the reduction of objective function value. The necessary time to reach for consensus (as described in section 2.2) is set as 100% of optimization time to assess the reduction of optimization duration using upper limit agreement criteria. Only combinations of adjustable parameters which reach consensus (Sulins and Mednis, 2012) are used in this study.

Upper limit agreement of two (three, four...) runs termination criterion is satisfied when within 1% consensus corridor of the best objective function value among all the parallel runs are two (three, four...) best values of objective function of parallel runs. There is no delay time applied (delay time=0).

.

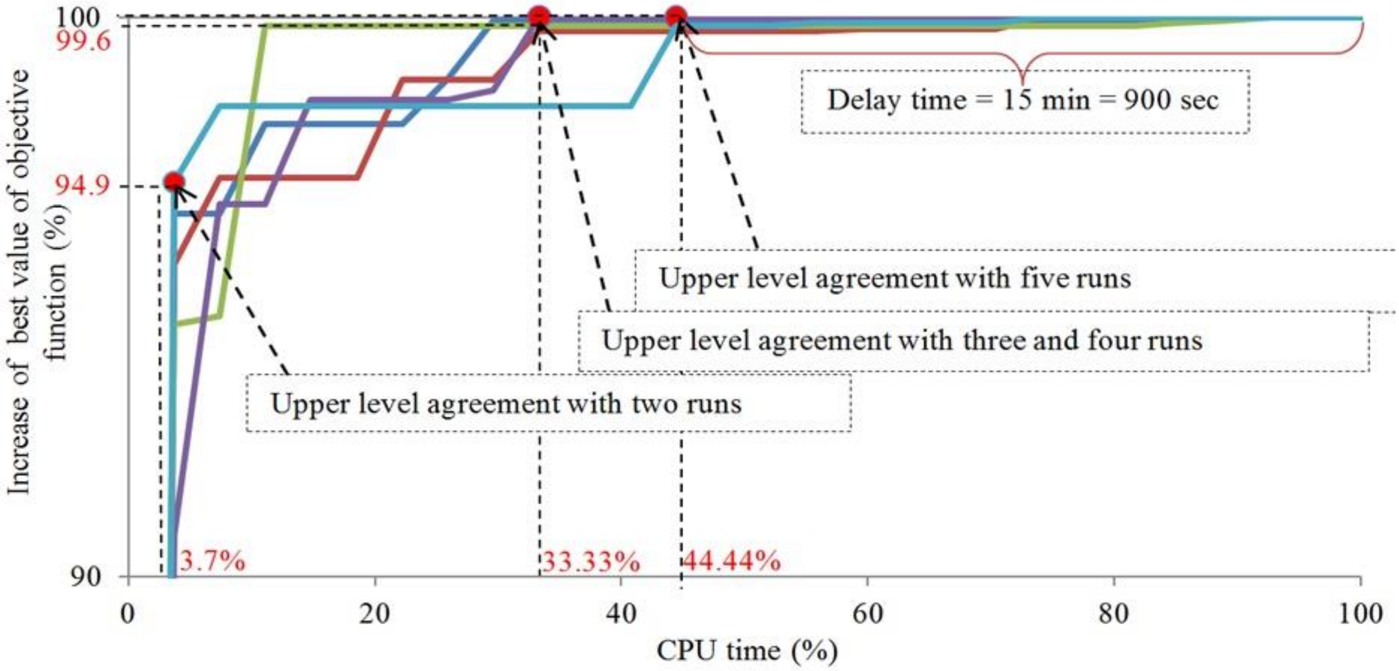

Fig. 1. **Determination of upper level agreement for two and more parallel optimization runs and corresponding losses of best value and savings of computational time. 0% and 100% of increase of objective function best value correspond to the value of model before optimization and after delay time of consensus correspondingly. 100% of optimization time correspond to 1620 seconds.**

3. Results

Earlier termination of parallel optimization runs reduces the duration of optimization and the best value of objective function (Fig. 2). Totally 63 combinations consist of correspondingly 6, 15, 20, 15, 6 and 1 combinations of 1, 2, 3, 4, 5 and 6 adjustable parameters in combination. The number of combinations that reached consensus/stagnation state are correspondingly 6/0 for one adjustable parameter in combination, 15/0 for two adjustable parameters in combination, 17/3 for three adjustable parameters in combination, 19/1 for four adjustable parameters in combination, 5/1 for five adjustable parameters in combination and 1/0 for six adjustable parameters in combination. Totally five out of 63 combinations stagnated and are not included in the calculations for Fig. 2. There is no statistics about the case of six adjustable parameters as there is only one combination possible.

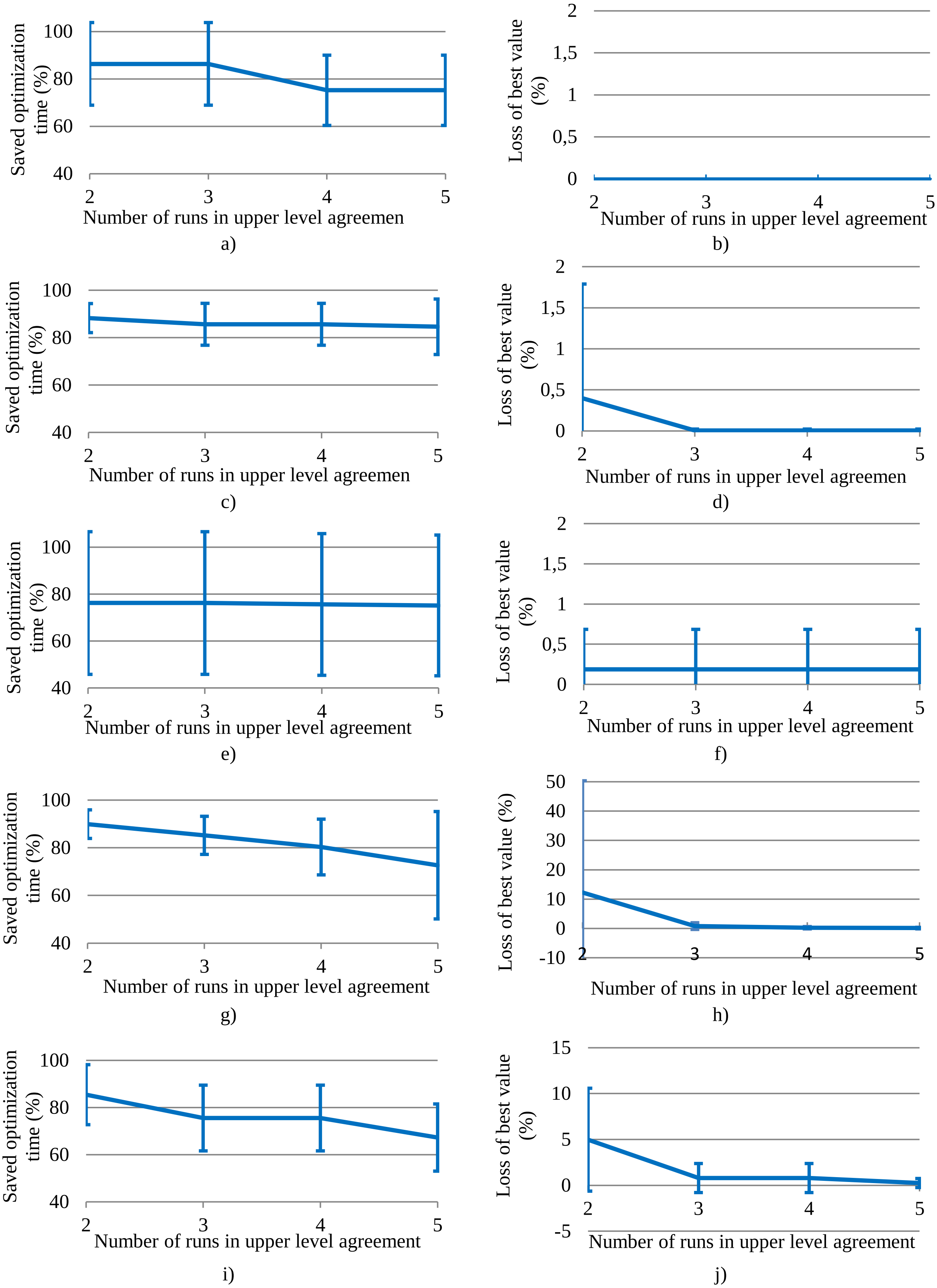

Fig. 2. **Reduction of objective function value and on the increase of optimization time for one (a and b), two (c and d), three (e and f), four (g and h) and five (i and j) parameters correspondingly. Error bars demonstrate the standard deviation at n=6 for one, n=15 for two, n=20 for three, n=15 for four and n=6 for five parameters.**

4. Discussion

The reduction of the optimization time compared with full consensus within 1% corridor and 15 minutes delay time is about 80% independent on the number of adjustable parameters per combination. The curves are decreasing but still the average values always remain above 60% indicating that the saving of optimization time can be expected about 60-90%. In absolute numbers the optimization duration savings increase with the number of parameters in combination.

The average loss of best value can be significant in case of upper level agreement of two runs while it is below 1% starting from upper level agreement of three runs. Thus experiments demonstrate that consensus criterion (Sulins and Mednis, 2012) can be applied in a less strict way saving about 2/3 of optimization time and loosing less than 1% of objective function improvement if upper level agreement of at least three parallel runs is reached.

More experiments are needed to generalize the conclusions about the efficiency of upper limit agreement criterion. In case bigger models and larger set of adjustable parameters the length of optimization duration would increase (Mozga and Stalidzans, 2011b) and time savings in percents would reduce while the time savings in absolute numbers would increase.

Looking at the savings of computational resources it can be calculated that faster result is reached by n-fold increase of computational resources where n is the number of parallel optimization runs. Thus in case of n-fold reduction of computational time would mean equal use of computational time (processor hours). More than n-fold reduction of computational time would mean additional benefit: savings of computational time in addition to the savings of optimization duration.

The 100% of optimization time in this study is determined experimentally and means the moment of automatic termination at consensus within 1% corridor of all parallel optimization runs. The time scale would therefore change if the corridor or delay time would be changed. The time scale would change even more if 100% time would be determined voluntary by an expert.

The algorithm for upper level agreement criteria implementation can be executed automatically in optimization software.

5. Conclusion

Significant time can be saved in case of approximate estimation of the best value of objective function for a particular combination of adjustable parameters using upper limit agreement criterion. Consensus of all parallel runs generally is a special case of upper level agreement criterion when all the parallel runs come to upper level agreement.

Computational experiments demonstrate that upper level agreement of at least three parallel optimization runs reduces the necessary optimization time by 60-90% and reduces the best value of objective function just by up to 1% compared to consensus within 1% of five parallel runs with delay time of 15 minutes. More extensive experiments would be needed to generalize this statement.

The optimization termination algorithm can be executed automatically.

References

Banga, J.R. (2008), "Optimization in computational systems biology.," *BMC systems biology*, Vol. 2, p. 47. http://dx.doi.org/10.1186/1752-0509-2-47

Galazzo, J. and Bailey, J.E. (1990), "Fermentation pathway kinetics and metabolic flux control in suspended and immobilized Saccharomyces cerevisiae," *Enzyme and microbial technology*, Vol. 12 No. 3, pp. 162–172. http://dx.doi.org/10.1016/0141-0229(90)90033-M

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., et al. (2006), "COPASI--a COmplex PAthway SImulator.," *Bioinformatics (Oxford, England)*, Vol. 22 No. 24, pp. 3067–74. http://dx.doi.org/10.1093/bioinformatics/btl485

Hübner, K., Sahle, S. and Kummer, U. (2011), "Applications and trends in systems biology in biochemistry.," *The FEBS journal*, Vol. 278 No. 16, pp. 2767–857. http://dx.doi.org/10.1111/j.1742-4658.2011.08217.x

Kostromins, A., Mozga, I. and Stalidzans, E. (2012), "ConvAn: a convergence analyzing tool for optimization of biochemical networks," *Biosystems*, Vol. 108 No. 1-3, pp. 73–77. http://dx.doi.org/10.1016/j.biosystems.2011.12.004

Mauch, K., Buziol, S., Schmid, J. and Reuss, M. (2001), "Computer-Aided Design of Metabolic Networks," *AIChE Symposium Series*, pp. 82–91.

Mendes, P. and Kell, D. (1998), "Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation," *Bioinformatics*, Vol. 14 No. 10, pp. 869–883. http://dx.doi.org/10.1093/bioinformatics/14.10.869

Mozga, I., Kostromins, A. and Stalidzans, E. (2011), "Forecast of Numerical Optimization Progress of Biochemical Networks," *Proceedings of the International Conference Engineering for Rural Development, 26-27 May 2011*, Jelgava, pp. 103–108.

Mozga, I. and Stalidzans, E. (2011a), "Convergence dynamics of biochemical pathway steady state stochastic global optimization," *Proceedings of IEEE 12th International Symposium on Computational Intelligence and Informatics (CINTI), 21-22 November 2011*, Budapest, IEEE, pp. 231–235. http://dx.doi.org/10.1109/CINTI.2011.6108504

Mozga, I. and Stalidzans, E. (2011b), "Convergence Dynamics of Biochemical Models to the Global Optimum," *Proceedings of the 3rd International Conference on E-Health and Bioengineering, 24-26 November 2011*, Iasi, pp. 227–230.

Nikolaev, E.V. (2010), "The elucidation of metabolic pathways and their improvements using stable optimization of large-scale kinetic models of cellular systems.," *Metabolic engineering*, Elsevier, Vol. 12 No. 1, pp. 26–38. http://dx.doi.org/10.1016/j.ymben.2009.08.010

Pentjuss, A., Odzina, I., Kostromins, A., Fell, D., Stalidzans, E. and Kalnenieks, U. (2013), "Biotechnological potential of respiring Zymomonas mobilis: a stoichiometric analysis of its central metabolism," *Journal of Biotechnology*, Vol. 165 No. 1, pp. 1–10. http://dx.doi.org/10.1016/j.jbiotec.2013.02.014

Rodríguez-Acosta, F., Regalado, C.M. and Torres, N.V. (1999), "Non-linear optimization of biotechnological processes by stochastic algorithms: Application to the maximization of the production rate of ethanol, glycerol and carbohydrates by Saccharomyces cerevisiae," *Journal of Biotechnology*, Vol. 68 No. 1, pp. 15–28. http://dx.doi.org/10.1016/S0168-1656(98)00178-3

Stalidzans, E., Kostromins, A. and Sulins, J. (2012), "Two stage optimization of biochemical pathways using parallel runs of global stochastic optimization methods," *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, IEEE, pp. 365–369. http://dx.doi.org/10.1109/CINTI.2012.6496792

Sulins, J. and Mednis, M. (2012), "Automatic termination of parallel optimization runs of stochastic global optimization methods in consensus or stagnation cases," *Biosystems and Information technology*, Vol. 1 No. 1, pp. 1–5. http://dx.doi.org/10.11592/bit.120501

Sulins, J. and Stalidzans, E. (2012), "Corunner: multiple optimization run manager for Copasi software," *Proceedings of 5th International Scientific Conference on Applied Information and Communication Technologies, 26-27 April 2012*, Jelgava, pp. 312–316.

Trinh, C.T. and Srienc, F. (2009), "Metabolic engineering of Escherichia coli for efficient conversion of glycerol to ethanol.," *Applied and environmental microbiology*, Vol. 75 No. 21, pp. 6696–705. http://dx.doi.org/10.1128/AEM.00670-09

Unrean, P., Trinh, C.T. and Srienc, F. (2010), "Rational design and construction of an efficient E. coli for production of diapolycopendioic acid.," *Metabolic engineering*, Elsevier, Vol. 12 No. 2, pp. 112–22. http://dx.doi.org/10.1016/j.ymben.2009.11.002