Report
T2002:15

ISRN:
SICS-T–2002/15-SE
ISSN: 1100-3154

GENFUNK

by

Erik Aurell, Mats Carlsson, Jan Ekman & Per Kreuger

16th September 2002

{eaurell,matsc,jan,piak}@sics.se
SICS, Swedish Institute of Computer Science
Box 1263, S-164 29 KISTA, SWEDEN

Abstract

This document summarizes the results obtained by SICS in project GENFUNK (2001). The project was carried out in collaboration with Global Genomics AB (Stockholm, Sweden). Jointly obtained results will be presented separately. Main funding was provided by Swedish Research Agency VINNOVA. Project GENFUNK studied a novel approach of measuring the global gene expression. In the method, mRNA is extracted from a tissue sample and transformed into cDNA captured on magnetic beads. This is then acted on by type IIS restriction endonucleases, which recognize certain short DNA sequences and cut the DNA close to those sequences. The resulting fragments are amplified in PCR with selected ligation fragments, and displayed in capillary electrophoresis. Determining the gene expression levels from the peak data is combinatorial optimization problem, which can in principle be solved, to give expression levels of most genes active in sampled cells, with good accuracy.

Keywords: Global gene expression, combinatorial assignment.

Document description: This document is structured as follows: The Introduction (Sect. 1) gives a general background, with an emphasis of the biological problem. Sect. 2 gives a more mathematical outline, and defines the terms used. Sect. 3 summarizes the experimental errors in the Global Genomics procedure that are pertinent for the algorithmic approach discussed here. Sect. 4 discusses an assignment problem, and Sect. 5 the choice of target functions. Sect. 6 discusses an integer programming model developed by Mats Carlsson. The implementation of this model, and tests, will be described elsewhere. Sect. 7 discusses a constraint programming approach developed by Per Kreuger and Jan Ekman. Sect. 8 briefly discusses a local search approach.

Document history: An earlier version, dated November 5, 2001, merged several documents, including text received from Sten Linnarsson, CTO of Global Genomics AB, on September 20, 2001. Further document history available from information in the CVS repository /home/sics/eaurell/CVSROOT.

Other documentation: The ideas and motivations behind Global Genomics' method are described in their patent [13]. A central technical point is the use of Type IIS restriction enzymes. The web site of the leading commercial provider, New England Biolabs, gives a wealth of information on these, and other, enzymes [16]. SICS and Global Genomics is preparing a joint paper, intended for *Proc. Nat'l. Acad. Sci. (USA)*, covering essentially Sect. 6 of this report, and in vivo and in silico tests [5]. Global Genomics and SICS are also preparing a second paper, intended for *Nature Biotechnology*, where the integrated procedure is compared with high-accuracy individual gene expression measurements (real-time PCR), and the today leading method for global gene expression measurements (microarrays) [14]. Adam Ameur and Jakub Orzekowski Westholm have expanded upon the ideas sketched below in Sect. 8. Their results are presented in their joint MSc thesis [1], and will perhaps be presented elsewhere. Jan Ekman (SICS) and Peter Lönnerberg (Global Genomics) are in charge of preparing a report on ideas how to optimally choose the restriction enzymes used in the Global Genomics procedure.

Acknowledgments: We thank VINNOVA for financial support under contract 341-2001-05001. E.A. thanks the Swedish Research Council for support under grant 621-2001-2704. We thank Lars Rasmusson for initiating this project from the side of SICS, and the management and staff of Global Genomics for generously sharing their ideas and technical details with us. We particularly thank Sten Linnarsson, Peter Lönnerberg and Mats Oldin for their time, for a pleasant and fruitful collaboration, and for patiently explaining the biological background and motivation.

1 Introduction

Defining the complexity that ensues when the relatively modest set of about 30 000 human genes is expressed has been called the next step in the genomic revolution [11, 26]. Preconditions for this step will be accurate, quick and cheap measurement techniques of what the cell is actually doing, on the levels of the transcriptome (read-outs of genes into mRNA), the proteome (proteins synthesized from mRNA templates), and the metabolome (complete set of metabolites, including proteins and other molecular species). Although progress has been made on the proteome, the most mature experimental techniques give access to the transcriptome. Even so, there is a clear trade-off between quantitatively precise methods, such as real-time PCR, which are labor-intensive, and methods well suitable to global analysis, such as microarrays, which have problems with accuracy and reproducibility.

In microarrays [23, 22] gene-specific probes are synthesized or deposited on a surface, which is then used to probe an mRNA sample. A major technical improvements was the use of direct oligonucleotide synthesis *in situ* using a mask-based light-directed system [7], later commercialized by Affymetrix as the GeneChip. Generally, microarrays have potential or actual problems of cross-talk, in that a reaction that goes for one probe may also, to some extent, go for another probe. Anecdotal evidence suggests substantial inter-array variations of measured mRNA abundances of the same sample. While these problems are likely to be overcome, the concomitant cost of using, for instance, quasi-redundant multiple probes for one gene may be significant. Microarrays finally also have the principal limitations of a closed system, in that they can only measure abundances with probes that have been chosen beforehand. For these reasons there have been several attempts to find alternatives to micro array technology, avoiding hybridization. Known examples in the literature include serial analysis of gene expression (SAGE) [25], integrated procedure for gene identification (IPGI) [27], introduce-amplified fragment length polymorphism (iAFLP) [10], GeneCalling [19], massively parallel signature sequencing (MPSS) [3], and total gene expression analysis (TOGA) [24].

There have also been attempts to quickly and accurately measure gene expression levels *partially*, a procedure known as gene expression profiling. Although this does not provide information about individual gene expression levels, the resulting profiles may nevertheless be highly specific to a given cellular state. A benchmark method has been differential display [12], where a cDNA population is separated by size on an electrophoresis gel. Brenner and Livak [4] used Type IIS restriction endonucleases to generate 5' overhangs of an unknown sequence, tagged the ends with fluorescent dyes, and separated these segements by length

on a polyacrylamide gel. Kato et al. [9] first divided a sample after cleavage by the enzyme into subpopulations, which were then separately amplified by PCR. This general approach using the special properties of Type IIS enzymes was subsequently taken up by several groups [21, 17, 15]. It is not directly suitable for direct gene expression analysis, since the information obtained usually is not sufficient to determine individual gene expression levels.

A recent proposal to turn Kato's method into a true global gene expression analysis tool is presented in [20], where an iterative procedure is used. The fragments after action by one enzyme and after ligation with an adapter, are then treated with a second restriction enzyme, the process repeated three times. In this set-up two base pairs are exposed at a time, giving six base pairs of information. The intended use of the method is however primarily genome indexing, not quantitative measurements of gene expression.

1.1 Statement of the problem

The key idea in the Global Genomics setup is to extend the methods of Brenner-Livak and Kato by dividing up a cDNA population into two or more subpopulations, and then use different restriction enzymes of the same type on each subpopulation. The technical steps in this procedure have not been fully published, except for a recently granted patent [13]. Nevertheless, it is quite clear and very likely that one can thus get much more information on the global gene expression levels, than by using only one restriction enzyme. It is also clear that the information is not directly available, but must be deduced from comparisons of the profiles from the respective restriction enzymes.

The goal of GENFUNK was to state precisely the above sketched combinatorial problem, and to develop methods to solve it. This document will center on one of the experimental set-ups used by Global Genomics, where three restriction enzymes are used. As will be described elsewhere [5], this allows for the determination of most expressed genes in mouse. In silico and laboratory experiments [14] further indicate that the method is quantitatively at least as accurate as microarrays, while still being an open system, capable of detecting any mRNA in the sample.

2 Mathematical modeling and definitions

Type IIS restriction enzymes, or interrupted palindrome restriction endonucleases, are proteins that recognize specific DNA base pair sequences, and cuts the DNA

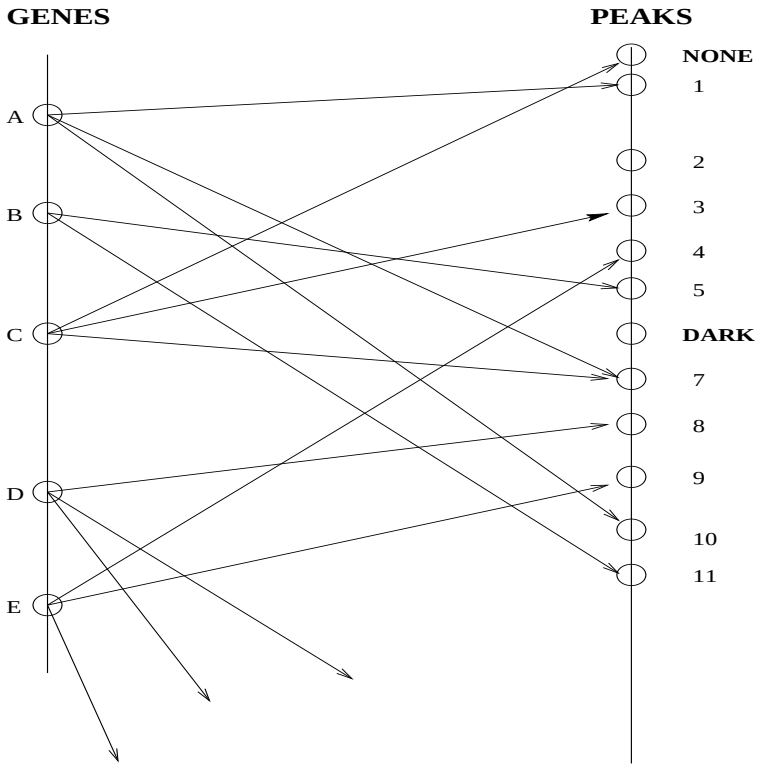

Figure 2.1: Bipartite graph illustrating the assignment problem of genes to peaks (or peaks to genes). The concepts genes, fragments and peaks are defined in the main text.

some distance away from the recognition site. The two strands are cut with a small offset, so the action leaves a DNA fragment with an overhang. The prime reference source for these enzymes is the New England Biolab web site [16]. For the present work one should preferentially choose enzymes with recognition sequences about five base pairs long, and that leave a four base pair overhang [13].

It is useful to represent the information of the identity of the recognition sequence as *enzyme*, or, more generally, as a *restriction group*. Global Genomics uses ligation fragments that on one side have a four base pair overhang, and primers in the PCR that on the other side have one of the three letters (G,T,C), followed by a poly-A string. That gives $3 \cdot 4^4 = 768$ possibilities. Each of these is referred to by Global Genomics as a *frame*. It might also be useful to have a concept of all the discrete information we have on fragments, frames and enzymes, e.g. a *frameset*. The setup considered here would hence have $3 \cdot 768 = 2304$ framesets. Each of these in fact labels one experiment, characterized by which restriction enzyme was used, which ligation fragments, and which primers in the PCR. Hence, for a particular species of mRNA to be likely to be observed, it should be present in at least about 2500 copies, so that it is present in every experiment, in particular in its own frameset. In practice, this means that the Global Genomics method can either measure highly expressed genes from one cell, or weakly expressed genes from a sample containing many cells. It is not a global gene measurement tool on the individual cell level.

A significant complication is the fact that many poly-adenylation sites are not precisely known. It is therefore convenient to define a *head-frame* as the four letters overhang, and a *sub-frame* as the single letter before the poly-adenylation sequence. The frame is the combination of the head-frame and the sub-frame.

Definition 2.1 *a gene is* ⟨**ID,s,stop,pal**⟩ *where* **ID** *is a unique identifier,* **s** *is a string of nucleotides,* **stop** *is a stop position and* **pal** *is a pointer to a list of alternative poly-adenylation sites. Such a list contains items* ⟨**ID,pos,q**⟩*, where* **ID** *is a unique identifier,* **pos** *is the position of the poly-adenylation site, and* **q** *is a quality indicator. We will later consider* **q** *indicators of the type of the uncertainty in* **pos***.*

Definition 2.2 *a fragment is* ⟨**enzyme,frame,length**⟩ *This represents a string of integer length* **length***, where we only know the information that it is produced by* **enzyme***, and occurs in* **frame***.*

Definition 2.3 *a peak is* ⟨**ID,enzyme,frame,length,Area**⟩ *where* **ID** *is an identifier provided with the processed experimental data.*

Definition 2.4 *a dark peak is* ⟨**dark,enzyme,frame,length,any**⟩. *This models a peak which was not observed because of technological limitations, e.g. machine failure, or because length too long or too short. It is not supposed to have been listed in the experimental data, but to have identifier* **dark**. *It can have any area.*

Definition 2.5 *a zero peak is* ⟨**zero, enzyme,frame,length,bg**⟩. *This models a peak which was not observed because its area was below the detection limit. It is not supposed to have been listed in the experimental data, but to have identifier* **zero**. *Its area must be less than* **bg** *(background).*

Description: There is a link from a gene to a peak if, for a given poly-adenylation site in pal, *i)* there is a recognition sequence in s for the $enzyme$ within about 1000 base pairs upstream from pos, *ii)* the overhang left by the restriction enzyme, is the same as in the *head-frame* of the peak, *iii)* there is a letter close to pos, up to the accuracy specified by q, that matches the *sub-frame* of the peak, *iv)* the distance from the recognition sequence closest pos is close to the $length$ of the peak. The problem is illustrated by the bipartite graph of Fig. 2.

As will be discussed in the next section, readings of the same gene with different poly-adenylation sites, are in fact quite analogous, to the Global Genomics algorithm, to separate genes. It is therefore convenient to introduce two more derived concepts:

Definition 2.6 *a* **t-gene**, *for* transcribed *or* terminated *gene, is* ⟨**ID,s,pos,Min,Max**⟩ *where* **ID** *is a unique identifier,* **s** *is a string of nucleotides,* **pos** *is the most likely position of the poly-adenylation site,* **Min** *is furthermost possible position of the poly-adenylation site in the 5' direction, and and* **Max** *the furthermost possible position of the poly-adenylation site in the 3' direction. A t-gene is formed from the information in one gene, and in one entry in its list of poly-adenylation sites. If there is no information in the quality indicator* **q** *on two hard bounds* **Min** *and* **Max**, *it can not form a t-gene.*

Definition 2.7 *A* **u-gene**, *for* unambiguous *gene, is a tuple* ⟨**ID,s,pos**⟩ *where* **pos** *is the* fixed *position of the poly-A site, and the other fields are the same as in a t-gene. Hence, the sub-frame of a u-gene is uniquely determined as* $\mathbf{s}[pos - 1]$.

3 Two sources of uncertainty

The bipartite graph encodes two sources of uncertainty.

3.1 Genes-Fragments uncertainty

The genomic information linking genes and fragments may be incomplete. One possibility is that the set of genes in itself is incorrect or incomplete, for (at least) the following reasons:

1. some genes may not be present in the data base at all;
2. there are alternative splicings of the same gene, and not all are in the data base;
3. there are alternative poly-A sites, and not all are in the data base;
4. for a given end, the sequence in the data base may not be complete all the way to the end

Two readings of the same gene, with the same poly-A site, but with different splicings, will produce the same 3' ends of mRNA. They can produce the same fragments, if the closest cleavage site by the restriction enzyme is where the two mRNA agree. They can also produce different fragments, if there is no cleavage site over the stretch where they agree, counting from the 3' end. These two strings of DNA (of different length) will however always agree at their 3' ends.

Two readings of the same gene, with the same splicing, but with different poly-A sites, will not produce the same 3' ends of mRNA. They will hence produce different fragments. They can produce strings of DNA that partially agree, over some of their length, counting from the ligation end, if there is no cleavage site between the two poly-A sites. If there is a cleavage site between the two poly-A sites, they will produce strings of DNA that typically do not agree at all. A gene, the true end which is uncertain, is similar in effect to an unknown alternative poly-A site.

A second possibility is sequencing errors, If those errors are at the bases specifying a frame, it would mean that a gene would consistently show up in the wrong frame. Alternatively, there are single nucleotide polymorphisms (SNPs), and a sample under consideration might come from an individual with another genotype than in the genomic data base. Celera quotes e.g. an average rate of SNP in the human of one per 1500 base pairs. There are ten base pairs in the frame and in the recognition sequence of the enzyme. If we assume $(1/1500)$ to be an error rate of a genomic sequence, the chance that all of ten base pairs are right is

$\left(1 - \frac{1}{1500}\right)^{10} \approx 0.94$. This rate is of course highly dependent on the accuracy of the genomic sequence, and on the rate of SNPs.

A third possibility is insufficient specificity of the ligation reaction. Ideally, two DNA sequences with complementary overhangs, say $\ldots \mathrm{TCGC}$ and $\mathrm{AGCG} \ldots$, should combine together, but $\ldots \mathrm{TCGC}$ should not combine with any other. In fact, the strands are held together by three hydrogen bonds between C and G, and two hydrogen bonds between A and T. There is hence some affinity also between, say, $\ldots \mathrm{TCGC}$ and $\mathrm{TGCG} \ldots$. One would therefore have to expect some cross-reaction. Global Genomics has data on these cross-reactions, and methods to circumvent the problem (information not public).

3.2 Fragments-Peaks uncertainty

The second source of uncertainty is that of experimental errors, in the Global Genomics setup, of the determination of $length$ and $Area$ of a peak, and in the comparison with fragments. We can assume that $Area$ is reproducible up to at least a factor two. True DNA string length is an integer, while the observed $length$ is approximately given with one decimal position [1].

Most of the scatter in $length$, of peaks, is in fact systematic, not random. One effect is that of translation from observational data to processed data, which is done with size markers. The passage time of a stretch of DNA through a capillary has a non-linear dependence on length. The mapping from passage time to length is performed by interpolation between the known size markers. There is a (length-dependent) error of this map, since there are only a finite number of size markers. Ideally, this should be corrected for.

Secondly, the mapping depends somewhat on the actual sequence. DNA may have a tendency to curl up, which varies with the sequence. That will effectively make pieces of DNA of the same length travel at different speeds in the capillary. This error may, as a first approximation, be modeled as a random spread. Nevertheless, it is not really random, because if one knows the letters in the DNA piece in question, and measures the passage of this sequence directly through the sequencing machine, one can determine the actual passage time. One can therefore envision that in a future version of this work, these length corrections will be known, and can be included as corrections in a data base of fragments.

[1] Resolution of the sequencing machine is about 11 data points per base, run-to-run variability specs are 15 % of one base.

4 Assignment

The actual experimental data from a tissue sample is a collection of peaks. Every peak can be the observation of fragments of one or several genes. The bipartite graph of Fig. 2 expresses all possible such observations; if a peak p can be an observation of a gene g there is a link (g, p). Let the set of links be denoted

$$\Sigma = \{(g,p); g \in \{Genes\}, p \in \{Peaks\}\} \tag{4.1}$$

It is convenient to consider more restricted link sets that go between *t-genes* and peaks:

$$\Sigma^t = \{(t,p); t \in \{\textit{t-Genes}\}, p \in \{Peaks\}\} \tag{4.2}$$

The difference between (4.1) and (4.2) is that a link in Σ only requires there is a suitable entry in the list of poly-adenylation sites, but does not define which one.

We recall (in this context) the definition of the *start* and the *end* of a link:

$$\forall\,(l = (g,p)) \in \Sigma: \quad start(l) = g \quad end(l) = p \tag{4.3}$$

We can therefore consider much information included in Σ (and in Σ^t):

$$\begin{array}{lcl} enzyme(l) & = & enzyme(end(l)) \\ length(l) & = & length(end(l)) \\ frame(l) & = & frame(end(l)) \end{array} \tag{4.4}$$

For $l \in \Sigma^t$ we have further information on which poly-adenylation site is considered. Enzyme $enzyme(l)$ cleaves the string s in t-gene t at a well-defined site, and leaves a fragment which agrees in frame with $frame(l)$. Its length is close to $length(l)$. We define the *predicted length* of link l to be the length of this fragment.

Definition 4.1 *Definition: an assignment is a subset σ of Σ^t such that for each t-gene tg, either there are no links in σ or there is a link from tg to one distinct peak per enzyme. The sub-frames of these peaks must be identical.*

An assignment can be considered the graph of a function $\sigma(tg, enzyme)$:

$$\sigma : \{(tg, enzyme) \to (p)\}, \quad \text{or } \sigma(tg,p) = p \text{ for } (tg,p) \in \Sigma^t \tag{4.5}$$

There may be additional acceptability conditions on assignments. These can be of the type that only some links from a given gene can be assigned together. They can also be of the type that the links that can be assigned from two genes vary together, e.g. that they have a common offset in length.

We can also extend the concept of assignment, to express that perhaps some genes are only expressed with a given poly-adenylation site in a given sample, or correlations between these expression level. Different poly-A sites of the same gene could then not be independently assigned to peaks. Similarly, one could also express that some genes would only appear with a given splicing in a given sample, or correlations between these expression levels.

5 Optimization

The goal is to determine gene expression levels. We introduce real non-negative variables

$$E_{tg} = \text{expression level of t-gene } tg \tag{5.1}$$

Given an assignment, we can then compute *predicted peak areas*:

$$E_p = \sum_{tg:p=\sigma(tg,enzyme(p))} E_{tg} \tag{5.2}$$

The predicted peak areas are functions of the assignment σ and the set of gene expression levels $\{E_{tg}\}$. Given an element in an assignment, we also have from above its predicted length $L_p(l)$.

We then have two kinds of errors. We can compare E_p with the measured peak areas, and L_p with the measured peak lengths.

5.1 A Bayesian detour

We will now make a short detour on target functions. Generally, one would often introduce some function $F(E_p, Area_p, L_p, length_p)$ to minimize. What does this mean? How to choose the functional form of F?

In Bayesian statistics this is framed as the following problem:

1. Let the DATA be the observed peak areas and peak lengths.
2. Let the MODEL be the assignment σ and the t-gene expression levels E_{tg}.
3. The measured lengths and areas may be considered realizations of random variables, where the distributions of the random variables encode the various error sources. These probabilities hence depend parametrically on σ and the E_{tg}'s, and can be written Prob$(\{Area\}, \{length\}; \sigma, \{E\})$. More compactly, these conditional probabilities are Prob$(\text{DATA}\,|\,\text{MODEL})$.

4. We seek to find the most likely model given the data, i.e. to maximize $\mathrm{Prob}(\mathrm{MODEL}|\mathrm{DATA})$. By Bayes' rule:

$$\mathrm{Prob}(\mathrm{MODEL}|\mathrm{DATA}) = \frac{\mathrm{Prob}(\mathrm{DATA}|\mathrm{MODEL})\,\mathrm{Prob}(\mathrm{MODEL})}{\mathrm{Prob}(\mathrm{DATA})} \tag{5.3}$$

   The denominator can be fixed by a normalization condition. The unconditional probability $\mathrm{Prob}(\mathrm{MODEL})$, Bayes' prior, is undetermined.

Different choices of Bayes' prior defines different best ways to solve the problem, given the knowledge, or belief, about the a priori likelihood of a given model, expressed by the prior. In each case, however, maximizing the probability in (5.3) is clearly equivalent to minimizing a target function

$$F = -\log \mathrm{Prob}(\mathrm{DATA}|\mathrm{MODEL}) - \log \mathrm{Prob}(\mathrm{MODEL}) \tag{5.4}$$

Every assumed target function can be given an interpretation in terms of assumed probability density functions, and vice versa. If there is no good reason to prefer one a priori distribution to another, or if the a priori distribution is indeed uniform over all models, then one will seek to just directly minimize the *maximum likelihood function:*

$$F^{ML} = -\log \mathrm{Prob}(\mathrm{DATA}|\mathrm{MODEL}) \tag{5.5}$$

The discussion of the rest of this section will be framed in the maximum likelihood language. Let us however remark that the a priori probabilities are most likely not uniform in the problem at hand. Indeed, it should be possible to determine reasonable a priori distribution by analysis of the data. The simplest ansatz is to take all gene expression levels a priori independent, but with a given distribution $p(E_{tg})$. The full target function would then be

$$F_{ind} = F^{ML} - \sum_{\text{t-genes}} \log p(E_{tg}) \tag{5.6}$$

Interestingly, we have then a compatibility condition, or a fixed point conditions; namely that the expression levels actually found from minimizing (5.6) should again be distributed as $p(E_{tg})$. With hundreds or thousands of expressed genes in a given sample this condition could be quite stringent. This point deserves further investigation.

5.2 All errors independent

Let us now consider just the maximum likelihood approach, and let first all measurement errors be independent random variables. Then, with some f and g,

$$\text{Prob}(Area, length; \sigma, E) = \prod_{\text{t-genes}} \prod_{\text{enzymes}} e^{-f(L_p, length_p)} \prod_{\text{peaks}} e^{-g(E_p, Area_p)} \tag{5.7}$$

and the target function is

$$F = \sum_{\text{t-genes}} \sum_{\text{enzymes}} f(L_p, length_p) + \sum_{\text{peaks}} g(E_p, Area_p) \tag{5.8}$$

One example is a quadratic target function for lengths

$$F_2 = \sum_{\text{t-genes}} \sum_{\text{enzymes}} \alpha_p (L_p - length_p)^2 \tag{5.9}$$

which is equivalent to assuming a *Gaussian* error distribution of lengths. Another example is an absolute value target function

$$F_1 = \sum_{\text{t-genes}} \sum_{\text{enzymes}} \alpha_p |L_p - length_p| \tag{5.10}$$

which is equivalent to assuming an *exponential* error distribution of lengths.

The peak areas are the outcomes of a multiplicative process (the PCR). Such processes often give rise to *log-normal distributions*. We may therefore also list a log-normal target function for areas

$$F_{ln} = \sum_{\text{peaks}} s_p \left(\log \frac{E_p}{Area_p} \right)^2 \tag{5.11}$$

Hard bounds on the length differences should be taken care of in the definition of Σ. Hard bounds on differences in expression levels may be taken care of by modifying the target function on areas.

5.3 Errors from one gene correlated

Proceeding further with the maximum likelihood method, one can also motivate slightly more complicated target functions. Suppose that the errors in length of the fragments produced from one gene with different enzymes are not independent.

That would be the case, for instance, if these fragments share DNA that is curled up, and travels faster in the capillary. With three enzymes, giving rise to peaks p_1, p_2, p_3 for a given gene, one would then have

$$F_{corr} = \sum_{\text{t-genes}} h(L_{p_1}, length_{p_1} \ldots, E_{p_1}, Area_{p_1} \ldots) \tag{5.12}$$

This formulation is particularly straight-forward for quadratic target functions. Let $\overrightarrow{\Delta l}$ be the vector of errors in length, with different enzymes, and M_{tg} a square matrix, then

$$F_{2,corr} = \sum_{\text{t-genes}} c_{tg} \left(\overrightarrow{\Delta l} \cdot M_{tg} \overrightarrow{\Delta l} \right) \tag{5.13}$$

This corresponds to a Gaussian probability distribution, where the mean of peak length in enzyme i is L_{p_i}, and the peak-peak length correlation is

$$\langle (length_{p_i} - L_{p_i})(length_{p_j} - L_{p_j}) \rangle = (M_{tg}^{-1})_{ij} \tag{5.14}$$

The matrix M_{tg} is hence the inverse of the correlation matrix.

6 A mixed-integer programming approach

In this section, we will develop a mixed-integer programming (MIP) model expressing the optimization problem. The model is closely related to the bipartite graph view; see Fig. 2. We now introduce the bipartite graph Σ^u where the nodes are u-genes and peaks. Σ^u is computed from Σ^t as follows:

1. Expand each t-gene (and its incident links) into multiple u-genes, one for each possible poly-A site.
2. Remove all links (ug, p) where $sub\!-\!frame(ug) \neq sub\!-\!frame(p)$.
3. Remove every u-gene (and its incident links) that does not have a link to one distinct peak per enzyme.
4. Remove all unconnected nodes.

The goal of the MIP model is to compute an assignment $\sigma \subseteq \Sigma^u$ and gene expression levels with minimal total cost; see below. In a first approximation, we use the measured peak areas as an upper bound on gene expression, and the peak area error is simply the total unaccounted for peak area. The mismatch in length between a u-gene and a peak gives the *penalty* of $(ug, p) \in \sigma$:

$$\rho(ug, p) : \textit{UGenes} \times \textit{Peaks} \rightarrow [0, 1] \tag{6.1}$$

For convenience, we define the *utility* of $(ug, p) \in \sigma$ as $1 - \rho(ug, p)$ The key idea of the MIP model is to capture the search space by decision (0-1) variables $B_{ug,p}$, one per link $(ug, p) \in \Sigma^u$. $B_{ug,p} = 1$ iff ug contributes to peak p. We also need to introduce an expression variable E_{ug} for every u-gene, and an ancillary variable $E_{ug,p}$ for every link. The model decomposes into independent subproblems, one per connected component of Σ^u.

6.1 Determinacy

A subproblem may easily be underdetermined, i.e. admit many optimal solutions that vary only in the expression levels of some groups of u-genes, reflecting a lack of data to resolve ambiguities. Rather than reporting an arbitrarily chosen solution, we would like to reflect the ambiguous evidence in the solutions reported. To this end, we define the equivalence relation:

$$ug_1 \sim_{\Sigma^u} ug_2 \stackrel{\textbf{def}}{=} \forall p \in \{\textit{Peaks}\} : (ug_1, p) \in \Sigma^u \Leftrightarrow (ug_2, p) \in \Sigma^u \tag{6.2}$$

and replace in the model each u-gene by the equivalence class to which it belongs, solve the MIP, and then infer bounds on the original expression variables from the solution found. In other words, we compute a reduced graph:

$$\hat{\Sigma}^u = \{(\hat{ug}, p) \mid \hat{ug} \in \{\textit{Classes}\}, p \in \{\textit{Peaks}\}\} \tag{6.3}$$

where each class $\hat{ug}$ is an equivalence class spawned by $\sim_{\Sigma^u}$, compute an assignment $\hat{\sigma} \subseteq \hat{\Sigma}^u$ and class expression levels with minimal total cost, and finally map those levels back to the original expression variables.

6.2 Variables

All variables are non-negative. Decision variables can take two possible values only, 0 and 1.

- A decision variable $B_{\hat{ug},p}$ for each link $(\hat{ug}, p) \in \Sigma^u$ such that $B_{\hat{ug},p} = 1$ iff $(\hat{ug}, p) \in \sigma$.
- An expression variable $E_{\hat{ug}}$ for each class.
- An expression variable $E_{cug,p}$ for each decision variable, the value of which is either 0 or equal to $E_{\hat{ug}}$.

- A slack variable S_p for each peak, denoting the part of it that is not assigned to any class.

6.3 Objective function

Within the MIP model, we can capture three relevant, linear criteria on a good solution: (i) maximal coverage, (ii) maximal expression, (iii) suppression of false positives. Specifically, we maximize a weighted sum $F(\sigma)$ of the following three terms wrt. the assignment σ: (i) negative total relative slack; (ii) total relative expression level weighted by utility; (iii) negative total penalty:

$$\begin{aligned} F(\hat{\sigma}) = \quad & -c_1 \textstyle\sum_p \frac{S_p}{Area(p)} \\ & +c_2 \textstyle\sum_{\hat{ug},p} \frac{(1-\hat{\rho}(\hat{ug},p))\cdot E_{\hat{ug},p}}{Area(p)} \\ & -c_3 \textstyle\sum_{\hat{ug},p} \hat{\rho}(\hat{ug},p) \cdot B_{\hat{ug},p} \end{aligned} \tag{6.4}$$

where $\hat{\rho}(\hat{ug},p) = \min_{ug\in\hat{ug}} \rho(ug,p)$. The coefficients c_1, c_2 and c_3 are subject to tuning.

6.4 Constraints

Valid assignments. The first type of constraints specify valid assignments. Every class $\hat{ug}$ can contribute to at most $|\hat{ug}|$ peaks per enzyme, where $|\hat{ug}|$ denotes the cardinality (number of elements) of $\hat{ug}$. Let M denote a very large constant.

$$\forall \hat{ug}, z : \sum_{p:enzyme(p)=z} B_{\hat{ug},p} \leq |\hat{ug}| \tag{6.5}$$

$$\forall \hat{ug}, p : E_{\hat{ug},p} \leq M \cdot B_{\hat{ug},p} \tag{6.6}$$

No peak overcoverage. The second type of constraints state that no peak should be overcovered:

$$\forall p : \sum_{\hat{ug}} E_{\hat{ug},p} + S_p = Area(p) \tag{6.7}$$

Consistent expression levels. The third type of constraints is a consistency relation between the expression levels of each class, and the ancillary variables $E_{\hat{ug},p}$ for each decision variable:

$$\forall \hat{ug}, z : E_{\hat{ug}} = \sum_{p:enzyme(p)=z} E_{\hat{ug},p} \tag{6.8}$$

Redundant constraints. Introducing a decision variable $B_{\hat{ug}}$ for each class, the following implied constraints can be added, subsuming constraint 6.5. It is unclear whether this would bring any computational benefit.

$$\forall \hat{ug}, z : \sum_{p:enzyme(p)=z} B_{\hat{ug},p} = B_{\hat{ug}} \tag{6.9}$$

$$\forall \hat{ug} : E_{\hat{ug}} \leq M \cdot B_{\hat{ug}} \tag{6.10}$$

6.5 Detecting u-gene and gene expression bounds

For each subproblem and each equivalence class $\hat{ug}$ it contains, the MIP solver provides *a)* the total expression, $E_{\hat{ug}}$, *b)* the set of peaks, $P(\hat{ug})$, that $\hat{ug}$ is assigned to *c)* for each $p \in P(\hat{ug}), E_{\hat{ug},p}$.

In addition, we know the individual u-genes making up $\hat{ug}$. Let $n = |\hat{ug}|$ and let $P_z(\hat{ug})$ denote $\{p \in P(\hat{ug}) \mid enzyme(p) = z\}$. We need an auxiliary array $a_z[i], 1 \leq i \leq n$, where:

$$\begin{gathered}\{0\} \cup \{a_z[i]\} = \{0\} \cup \{E_{\hat{ug},p} \mid p \in P_z(\hat{ug})\} \\ a_z[1] \leq \cdots \leq a_z[n]\end{gathered} \tag{6.11}$$

Lower and upper bounds for $E_{ug}, ug \in \hat{ug}$ are obtained using rule 6.12, which is valid for any enzyme z:

$$a_z[1] \leq E_{ug} \leq a_z[n] \tag{6.12}$$

Also, for each gene g, we know the number $nug(g, \hat{ug})$ of its u-genes that are included in $\hat{ug}$. Let $E_g^{\hat{ug}}$ denote the contribution of the class to the expression of g. Lower and upper bounds for $E_g^{\hat{ug}}$ where $nug(g, \hat{ug}) > 0$ are obtained using rule 6.13, which is valid for any enzyme z. Note that these bounds are much sharper than those obtained by rule 6.12.

$$\sum_{i=1}^{nug(g,\hat{ug})} a_z[i] \leq E_g^{\hat{ug}} \leq \sum_{i=n-nug(g,\hat{ug})+1}^{n} a_z[i] \tag{6.13}$$

Finally, to obtain lower and upper bounds for E_g, the expression of each gene g, we simply compute the sum of the contributions:

$$\sum_{\hat{ug}} \min(E_g^{\hat{ug}}) \leq E_g \leq \sum_{\hat{ug}} \max(E_g^{\hat{ug}}) \tag{6.14}$$

6.6 Final remarks

Further constraints on valid assignments can be readily expressed as constraints on the $B_{\hat{ug},p}$ variables.

Many MIP solvers handle so called SOS Type 1 constraints such as constraint (6.5) specially. CPLEX, for example, uses special branching strategies for such sets of variables, and lets the user provide weight information to guide the search. The weight for $B_{\hat{ug},p}$ could e.g. be chosen as $\hat{\rho}(\hat{ug}, p)^{-1}$.

7 A constraint programming approach

Constraint programming (CP) is a general technique for declarative description and effective solving of large combinatorial problems, with particular applications to planning and scheduling [2, 8].

The idea of constraint programming is that the user solves problems by stating constraints (conditions, properties) which must be satisfied by the solution. A computation engine then executes a search algorithm to find a solution. A good general constraint solver, e.g. [6], allows for both the user explicitly specifying the search procedures, and for these being chosen by the solver.

Constraint programming can from one point of view be seen as a user interface to vanilla-flavor satisfiability and optimization problems. The capability of a solver hence depends both on the quality of the translation of the constraints (provided by the user) into a search/optimization problem, and on the power of the algorithms the solver then employs. The following discussion will be geared to a formulation where the search algorithms are efficient on linear constraints. It therefore assumes an underlying computational efficiency comparable to the mixed integer programming approach of Sect. 6. One advantage of constraint programming is that the formulation is closer to the original problem, and frees the user from the obligation of casting the same in a format suitable to the search/optimization procedure. This aspect will be evident in the following description. We note that a partial implementation has been carried out in the project, but not a full implementation that would allow comparison with the mixed integer programming approach of Sect. 6.

7.1 Notations used

The predicted lengths of all the t-genes under the enzymes can be represented as a $n \times m$ matrix

$$L = \begin{matrix} L_{11} & \cdots & L_{n1} \\ \vdots & \ddots & \vdots \\ L_{1m} & \cdots & L_{nm} \end{matrix}$$

Here n is the number of t-genes, m is three in the standard present setup, and each element L_{ij} represents the expected length of the fragment of the i'th gene, from the 3' end to the closest recognition sequence in the j'th RESTRICTION GROUP[2]. The *predicted head frame* of any t-gene is the overhang left by an enzyme that cleaves a given sequence at the recognition site closest to pos. It is convenient to define this as a function:

Definition 7.1 *Let ϕ be a function from a t-gene i and an enzyme j to the initial sequence of the fragment produced from i by j. Then $k = \phi(i, j)$ is the predicted head frame of t-gene i under enzyme j.*

Given a value of an offset we can also compute the *predicted sub-frame* of a t-gene. This simply means that if the t-gene would end at position O, which lies within o of pos, then the predicted sub-frame is the letter at O. It is also convenient to give this as a function, viz.

Definition 7.2 *Let ς be a function from a t-gene i and a (positive or negative) offset o from the expected length given by L_{ij} to a sub-frame s such that* if *the actual polyadenylation site would give a fragment of length of $L_{ij}+o$ in restriction group j,* then *the expression of t-gene i for that restriction group will occur in sub-frame $\varsigma\,(i, o)$ (of head-frame $\phi\,(i, j)$).*

Definition 7.3 *Let the* RESULT OF AN EXPERIMENT *be represented by a finite list of peak parameters*

$$P = P_1, \ldots, P_p$$

sorted by ascending peak position and where the values of each parameter is referred to as follows:

Let τ represent the base pair resolution of the experimental equipment, which can currently be rounded to 11 steps per base pair, and let $\pi_l = \pi\,(l)$ denote the observed position of an observed peak given in units of $1/\tau$:th of a base pair, $\alpha_l = \alpha\,(l)$ the observed area in arbitrary units and let furthermore $\gamma_l = \gamma\,(l)$ denote the restriction group, $\phi_l = \phi\,(l)$ the head-frame and $\varsigma_l = \varsigma\,(l)$ sub-frame

[2]In the standard current setup, the j'th out of $m = 3$ restriction endonucleases.

of each peak P_l[3]. Since P is sorted by peak position, we know that the following condition holds

$$\forall kl\,(k < l \Rightarrow \pi\,(k) \leq \pi\,(l)) \tag{7.1}$$

7.2 Matchings

The central problem is that several t-genes may contribute to a particular peak. With the simplifying conjecture that a t-gene can contribute to at most one fragment length in each restriction group, also made elsewhere in this report, we can represent a matching between t-gene expressions and observed peaks as an assignment of peak indexes to t-genes. Because of the uncertainty of the exact polyadenylation site, we need to handle regular offsets occurring in all restriction groups.

Definition 7.4 *Let a* MATCHING *be represented by an assignment for each t-gene i of a (discrete) offset variable O_i and of m (discrete) peak variables $C_{i1}, \ldots, C_{im}$ such that $C_{ij} = 0$ if the t-gene does not contribute to any peak in a particular restriction group j and $C_{ij} = l$ for some $0 < l \leq p$ otherwise*[4].

We require any matching to fulfill for each t-gene i and restriction group j the following conditions:

$$\forall\,(i \leq n)\,\forall\,(j \leq m)\,(C_{ij} \neq 0 \Rightarrow \gamma\,(C_{ij}) = j) \tag{7.2}$$

In other words, any assigned peak occurs in the correct restriction group. The peak also has to occur in the correct head-frame and for a given offset O_i the correct sub-frame, i.e.:

$$\forall\,(i \leq n)\,\forall\,(j \leq m)\,(C_{ij} \neq 0 \Rightarrow \phi\,(C_{ij}) = \phi\,(i,j)) \tag{7.3}$$

and

$$\forall\,(i \leq n)\,\forall\,(j \leq m)\,(C_{ij} \neq 0 \Rightarrow \varsigma\,(C_{ij}) = \varsigma\,(i,O_i)) \tag{7.4}$$

Let each variable O_i have initial domain constrained by

$$\forall\,(i \leq n)\,(\lambda_i \leq O_i \leq \upsilon_i) \tag{7.5}$$

[3]Note that π and γ correspond to the $length$ and $enzyme$ function of Sect. 4 and that the $frame$ function of that section is here split into the two functions ϕ and ς.

[4]The notion of matching defined here is very close to that of Sect. 4. We have here one variable C_{ij} for each t-gene i and restriction group j. Assigning a value l to one of these variables corresponds exactly to letting the value of the expression $\sigma\,(i,j)$ be l.

where λ_i and υ_i impose hard limits on the offset from the expected polyadenylation site for a given t-gene i. To capture the fact that the expression of a t-gene cannot occur in one restriction group and not in another (unless the expected fragment length for that group is 0) we enforce the following condition on the matching variables for each t-gene i.

$$\forall\,(i \leq n)\,(\exists j\,(C_{ij} = 0 \wedge L_{ij} \neq 0) \Rightarrow (O_i = 0 \wedge \forall\,(j \leq m)\,(C_{ij} = 0))) \quad (7.6)$$

7.3 Matching errors and penalty

The O_i variable captures the uncertainty in the exact position of the polyadenylation site while the possible assignments for the C_{ij} variables should depend on the position π_l of each candidate peak l. We will define an error variable, which we will use to enforce further necessary conditions on the matching and a matching penalty.

Definition 7.5 *Let the* MISMATCH ERROR Me_{ij} *of a particular t-gene i in a restriction group j be defined as*

$$Me_{ij} = \begin{cases} 0 & if \quad C_{ij} = 0 \\ \tau L_{ij} + \tau O_i - \pi\,(C_{ij}) & otherwise \end{cases} \quad (7.7)$$

Since Me_{ij} encode the absolute offset from the expected fragment length we can use it to constrain the matching variables as follows. Let each C_{ij} have an initial domain $0, \ldots, p$ and constrain the C_{ij} and O_i variables by enforcing for each t-gene i and restriction group j the following condition

$$\lambda_{ij} \leq Me_{ij} \leq \upsilon_{ij} \quad (7.8)$$

where λ_{ij} and υ_{ij} are parameters encoding the maximum acceptable absolute error in detected fragment length in an experiment for t-gene i in restriction group j.

Motivated by these observations we will enforce limits on the discrepancies between the errors in the restriction groups for any given t-gene i: For each t-gene i and any two restriction groups j and k such that $C_{ij} \neq 0$ and $C_{ik} \neq 0$ enforce a condition of the following form

$$|Me_{ij} - Me_{ik}| < \delta\,(|L_{ij} - L_{ik}|) \quad (7.9)$$

where δ encode the acceptable level of error discrepancy as function of the difference in fragment length. Typically, the maximum difference in error should vary from 1τ or 2τ to maybe 5τ or 10τ.

We can also state a limit on the error contribution for a given offset O_i from the expected fragment length for a particular t-gene i:

$$\sum_{j \leq m \wedge C_{ij} \neq 0} |Me_{ij}| < \mu_i \tag{7.10}$$

where $\mu_i$represents a maximum acceptable total error over the m restriction groups for t-gene i.

Generally, the above conditions do not uniquely determine the assignment variables and it is an optimization problem to compute the best assignment for a particular experiment and some given cost function. Part of such a cost function could be a weighted sum of the above errors:

Definition 7.6 *Let the* MATCHING PENALTY *for a given matching of a particular t-gene i be defined by the following expression:*

$$Mp_i = \sum_{j \leq m \wedge C_{ij} \neq 0} \mu_{ij} \left(Me_{ij}\right)_i + \sum_{j=1}^{m-1} \left(\sum_{k=j}^{m} \delta_{ijk} \left(Me_{ij} - Me_{ik}\right) \right) + \kappa_i \left|O_i\right|$$

where μ_{ij} are function parameters encoding the relative weight we give to each individual match error for a given t-gene i and restriction group j, δ_{ijk} are constant parameters encoding the relative weight given to error discrepancies for each pair $\langle j, k \rangle$ of restriction groups and κ_i is a constant parameter encoding the relative penalty, and κ_i is a constant parameter encoding the relative penalty given to offsets in polyadenylation site for t-gene i.

The μ_{ij} can in an assumed underlying essentially linear optimizer still depend both on the sign and the size of $Me_{ij} - O_i$. We here assume such a dependence, and also a dependence on the expected fragment length in each enzyme group. Similarly, the δ_{ijk} should probably depend on the size and difference in expected fragment lengths for the gene i in enzyme groups j and k. The κ_i should probably depend on the confidence we put the polyadenylation site used to compute the expected fragment lengths for gene i. We note that the matching penalty (Def. 7.6), for all its complexity, still does not take into account any quantitative measures at all. In principle we can make an assignment of 0 to all C_{ij} variables that give a penalty of 0 and fulfill all necessary condition mentioned so far. In order to really assess an assignment we need to take into account also the quantitative expression of each t-gene and the areas of peaks assigned to the t-gene.

7.4 Quantitative comparison with expression levels

Definition 7.7 *Let the quantitative level of expression of the i:th t-gene be represented by a variable E_i in the same unit as that used to express the area α_l of each peak P_l*[5]*.*

We make the following observations:

1. The sum of contributions to a particular observed peak may not (up to the some error) exceed the area of the peak itself
2. The opposite is not true since there may be unknown t-genes contributing to any particular observed peak
3. The matching procedure should however aim to minimize the sum of such "unexplained" peak areas

Based on the first observation we can define for each peak in an experiment a necessary condition:

Condition 7.8 *We enforce for each observed peak P_l and any given matching the following condition on the expression variables:*

$$\forall\,(l \leq p)\left(\left(\sum_{\left\{i|C_{i\gamma(l)}=l\right\}} E_i\right) \leq \zeta_l \alpha\,(l)\right) \tag{7.11}$$

where ζ_l is a parameter representing a maximum acceptable AREA UNDERESTIMATION *of the area of the peak l. We let ζ_l depend on the position π_l of peak P_l.*

A value of 2 for ζ allows a factor two of area underestimation of a peak in the experiment, which appears to be a realistic first estimation. This formulation does not take care of systematic errors such as those discussed above in Sect. 3, which we assumed have been corrected for. To be able to penalize both area underestimation and unexplained area of observed peaks we will define first an absolute error and then based on this quantitative penalties for a given matching.

Definition 7.9 *Let the* ABSOLUTE QUANTITATIVE ERROR Qe_l *for each observed peak l and any given matching be defined by*

$$Ae_l = \alpha\,(l) - \left(\sum_{\left\{i|C_{i\gamma(l)}=l\right\}} E_i\right) \tag{7.12}$$

[5]The E_i variables correspond exactly to the E_{tg} variables of Sect. 5

Note that if there is no t-gene assigned to a particular peak the absolute quantitative expression error will be equal to the area of the peak. Note also that in the case of area underestimation the absolute error will be negative.

Definition 7.10 *Let the area underestimation penalty Ap_l for a particular peak P_l be defined by the following expression*

$$Ap_l = \begin{cases} Ae_l & if \quad Ae_l \geq 0 \\ 0 & otherwise \end{cases} \tag{7.13}$$

Definition 7.11 *Let the area underestimation penalty Up_l for a particular peak P_l be defined by the following expression*

$$Up_l = \begin{cases} \frac{-Ae_l}{\alpha_l} & if \quad Ae_l < 0 \\ 0 & otherwise \end{cases} \tag{7.14}$$

We will now define a cost function to replace the one suggested in Def. 7.6 above with one based on the quantitative expression error as follows:

Definition 7.12 *Let the* TOTAL PENALTY *for a given matching and a given assignment of the expression variables E_i be defined by the following expression:*

$$\sum_{i \leq n} Mp_i + \varpi \sum_{l=1}^{p} Ap_l + \omega \sum_{l=1}^{p} Up_l \tag{7.15}$$

where ϖ and ω are some suitable weights expressing the relative contribution to the penalty from area underestimation and unexplained area of observed peaks respectively

Assign values to E_i and discrete values to C_{ij} and O_i for all $0 < i \leq n$ and $0 < j \leq m$ so as to minimize the expression in Def. 7.12 subject to the conditions on matchings expressed in equations (7.2), (7.3), (7.4) (7.5) and (7.6), the bounds on matching errors expressed in equations (7.8), (7.9) and (7.10), and condition (7.8) on the area underestimation.

7.5 Final notes on search procedure

For purposes of search we can decompose the problem into independent subproblems as follows.

Definition 7.13 *Let for each t-gene i its* CANDIDATE PEAKS *in each restriction group be the domain of the variable C_{ij} constrained by the conditions above.*

Definition 7.14 *Let furthermore the* CANDIDATE t-GENES *of a peak P_l be all the t-genes whose candidate peaks include P_l.*

The set of candidate t-genes of the candidate peaks of a particular t-gene can extended by recursively considering candidate t-genes and peaks.

Definition 7.15 *Let the cluster of a particular t-gene be the transitive closure of the candidate t-genes of its candidate peaks.*

The utility of an assignment of all variables associated with a particular t-gene can be assessed by considering its relative contribution to the the areas of peaks assigned to it. This could be used to formulate heuristics for search of optimal assignments.

Definition 7.16 *Let the relative contribution of a t-gene i to the peaks it is assigned to be defined by the following expression*

$$\sum_{j=1}^{m} \begin{cases} \frac{E_i}{\alpha(C_{ij})} & if \quad C_{ij} \neq 0 \\ 0 & otherwise \end{cases} \tag{7.16}$$

8 Local search

Local search is a general designation of heuristics that try to find an optimum of a nonlinear minimization problem by small local steps. In a way, they generalize methods of the continuum that try to go down a gradient, viz. the conjugate gradient method [18], and share the same defect: in general it is difficult to know if one has found a global minimum, or just a local minimum. In practice one therefore often restarts the method more than once, and checks a posteriori if the results of this sampling are consistent with a true minimum been found.

Three examples of often used methods of local search are *genetic algorithms*, *simulated annealing* and *taboo search*. Details and an implementation of a meta-algorithm using complete enumeration, taboo search and best/first improvements on subproblems of different sizes are presented elsewhere [1]. We will here briefly describe simulated annealing in the present context.

We want to define *state variables* and *state space* of the problem. Let out(tg) be the set of links in Σ^t that start at t-gene tg. Associate to each of these links l a variable s_l, equal to 0 or 1. We can refer to this variable as the *spin* of the link; it is analogous to the decision variables $B_{ug,p}$ in Sect. 6. The state variables are the link spins. As auxiliary variables, we have the gene expression levels of (5.1).

A set of link spins is identified with an assignment σ according to

$$\forall l = (tg, p) \in \Sigma^t \quad ((s_l = 1) \equiv (l \in \sigma)) \tag{8.1}$$

The state space is hence the set of configurations of the link spins that can be identified with an assignment. The auxiliary gene expression variables must be non-negative. Following the discussion in Sect. 5 we can assume an a priori probability of a configuration of the state variables and the auxiliary variables. We now assume that we can solve for the auxiliary variables $\{E\}$ that maximizes probability for a given configuration of the spin variables $\{s\}$.

That will then give a probability of a configuration of the spin variables only. In statistical mechanics, such a probability would be proportional to $e^{-\frac{\mathcal{E}(\{s\})}{T}}$, where $\mathcal{E}$ is the *energy* of the configuration, and T the *absolute temperature*. Up to the temperature, the energy can be identified with the target function F of (5.4). We introduce the concepts *state sum* and *free energy*

$$\mathcal{Z}(T) = \sum_{\{s\}} e^{-\frac{\mathcal{E}}{T}} \qquad \mathcal{F}(T) = -T \log \mathcal{Z}(T) \tag{8.2}$$

We have, at least in systems with a unique global minimum, that the free energy at zero temperature is the minimum energy

$$\text{Min}\, \mathcal{E} = \text{Lim}_{T \to 0}\, \mathcal{F}(T) \tag{8.3}$$

Equations (8.2) and (8.3) form the basis for simulated annealing. The procedure is as follows:

- start with some value T
- start with some initial configuration $\{s\}$
- generate local changes to $\{s\}$.
- always accept the change if they lower energy
- accept the change with probability $e^{-\frac{\Delta\mathcal{E}}{T}}$ if the energy change $\Delta\mathcal{E}$ is positive
- generate enough changes such that the average free energy at this temperature has stabilized
- lower temperature and loop

The major problems in implementing a simulated annealing procedure is how to generate the local moves, so that they cover all of state space, and how to lower temperature. Both may require trial-and-error experiments [18].

9 Conclusion

The first and main message is that the invention of Global Genomics naturally leads to a combinatorial optimization problem. With reasonable choices of the target function, such that it is finitary and linear, this problem can efficiently be solved with mixed-integer programming (MIP) techniques. Such techniques have the great advantage, in an industrial setting, that the limitations of the method are well understood, and that very good commercial solvers are available. We believe in fact that these may have much wider applicability in Bioinformatics.

A second message is that, as an alternative to MIP one may also use constraint programming (CP) techniques. Although we have not checked that on this problem, CP can in principle handle more complex linear constraint structures than MIP, and can thus give an advantage in taking some classes of experimental errors into account. We have in this project performed a feasibility study, and implemented some parts of the CP solver.

The third and final message is that heuristics, local search techniques, can also be used, and may allow even better modeling of the error sources. In a continuation of this work, a local search solver has been implemented, and compared with the MIP solver. All three models can, and should, be evaluated against quantitatively precise direct measurements of expression levels on a gene-by-gene basis. At present, such tests are only in the pipeline for the fully operational MIP solver.

References

[1] Adam Ameur and Jakub Orzechowski Westholm. Local search methods in gene expression analysis. Uppsala Master's Theses in Computing Science 222, 2002. ISSN 1100-1836.

[2] Roman Bártak. *Guide to Constraint Programming*. Charles' University, Prague, Czech Republic, 2002. http://kti.ms.mff.cuni.cz/ bartak/constraints/.

[3] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18:630–634, 2000.

[4] S. Brenner and K.J. Livak. DNA fingerprinting by sampled sequencing. *PNAS*, 86:8902–06, 1989.

[5] Mats Carlsson, Erik Aurell, Jan Ekman, Per Kreuger, Sten Linnarsson, Peter Lönnerberg, and Mats Oldin. Global gene expression analysis by combinatorial peak assignment. [in preparation, to be submitted to PNAS], 2002.

[6] Mats Carlsson et al. *SICStus Prolog User's Manual*. Swedish Institute of Computer Science, release 3 edition, 1995. http://www.sics.se/sicstus.

[7] S.P. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251:767–773, 1991.

[8] Michael Jampel, David Joslin, and Peg Eaton. *Constraints Archive*. University of New Hampshire, 2002. http://kti.ms.mff.cuni.cz/ bartak/constraints/.

[9] K. Kato. Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Research*, 23:3685–3690, 1995.

[10] S. Kawamoto, T. Ohnishi, H. Kia, and O. Chisaka. Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. *Genome Research*, 9:1305, 1999.

[11] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[12] P. Liang and A.B. Pardee. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257:967–971, 1992.

[13] S. Linnarsson, P. Ernfors, and G. Bauren. Methods for analysis and identification of transcribed genes, and fingerprinting. World Intellectual Property Organisation, International Publication Number 02/08461 A2, 2002.

[14] Sten Linnarsson and *et al*. Tangerine (draft designation). to be submitted to Nature Biotech, 2002.

[15] H. Mahdeva, M.P. Starkey, F.N. Sheikh, C.R. Mundy, and N.J. Samani. A simple and efficient method for the isolation of differentially expressed genes. *Journal of Molecular Biology*, 284:1391–1398, 1998.

[16] New England Biolabs Inc. *REBASE*, 2002. http://rebase.neb.com/rebase/rebase.html.

[17] Y. Prashar and S.M. Weissman. Analysis of differential gene expression by display of 3' restriction fragments of cDNA. *PNAS*, 93:659–663, 1996.

[18] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes*. Cambridge University Press, 1988.

[19] R.A. Shimkets, D.G. Lowe, J.T. Tai, P. Sehl, H. Jin, R. Yang, P.F. Predki, B.E. Rothberg, M.T. Murtha, M.E. Roth, et al. Gene expression analysis by transcript profiling coupled to a gene database query. *Nature Biotechnology*, 17:798–803, 1999.

[20] D.R. Sibson and F.E.M. Gibbs. Molecular indexing of human genomic DNA. *Nucleic Acids Research*, 29:e95, 2001.

[21] D.R. Sibson and M.P. Starkey. *Increasing the average abundance of low abundance cDNAs by ordered division of cDNA populations*, volume 69, pages 13–32. Humana Press Inc., 1997.

[22] E.M. Southern and U. Maskos. Parallel synthesis and analysis of large numbers of related chemical compounds: applications to oligonucleotides. *J Biotechnol*, 35:217–227, 1994.

[23] E.M. Southern, U. Maskos, and J.K. Elder. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics*, 13:1008–1017, 1992.

[24] J.G. Sutcliffe, P.E. Foye, M.G. Erlander, B.S. Hilbush, L.J. Bodzin, J.T. Durham, and K.W. Hasel. TOGA: an automated parsing technology for analyzing expression of nearly all genes. *PNAS*, 97:1976–1981, 2000.

[25] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[26] C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[27] S.M. Wang and J.D. Rowley. A strategy for genome-wide gene analysis: Integrated procedure for gene identification. *PNAS*, 95:11909–11914, 1998.