PA 1014547

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

May 23, 2003

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: ***60/382,628***
FILING DATE: ***May 24, 2002***

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

T. Wallace
T. WALLACE
Certifying Officer

A/PROV

PTO/SB/16 (6-9

Approved for use through 04/11/98. OMB 0651-00

Patent and Trademark Office, U.S. DEPARTMENT OF COMMERC

PROVISIONAL APPLICATION COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION under 37 CFR 1.53 (c).

Docket Number	620-200	Type a plus sign (+) inside this box→

INVENTOR(S)/APPLICANT(S)

LAST NAME	FIRST NAME	MIDDLE INITIAL	RESIDENCE (CITY AND EITHER STATE OR FOREIGN COUNTRY)
LINNARSSON	Sten		Global Genomics AB, Berzelius vag 3, S-171 77 Stockholm, Sweden
LÖNNERGERG	Peter		Global Genomics AB, Berzelius vag 3, S-171 77 Stockholm, Sweden
OLDIN	Mats		Global Genomics AB, Berzelius vag 3, S-171 77 Stockholm, Sweden
AURELL	Erik		Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden
CARLSSON	Matt		Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden
EKMAN	Jan		Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden
KREUGER	Per		Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden
RASMUSSON	Lars		Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden

TITLE OF THE INVENTION (280 characters)

PROFILING MOLECULES

CORRESPONDENCE ADDRESS

Mary J. Wilson
NIXON & VANDERHYE P.C.
1100 North Glebe Road
8th Floor
Arlington

STATE	Virginia	ZIP CODE	22201	COUNTRY	U.S.A.

ENCLOSED APPLICATION PARTS *(check all that apply)*

☒	Specification	Number of Pages	66	☐	Applicant claims "small entity" status. ☐ "Small entity" statement attached.
☒	Drawing(s)	Number of Sheets	3	☐	Other (specify)

METHOD OF PAYMENT *(check one)*

☒	A check or money order is enclosed to cover the Provisional filing fees ($160.00)/($80.00)	PROVISIONAL FILING FEE AMOUNT ($)	160.00
☐	The commissioner is hereby authorized to charge filing fees and credit Deposit Account Number 14-1140		

The invention was made by an agency of the United States Government or under a contact with an agency of the United States Government.

☒ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are:

Respectfully submitted,
SIGNATURE Mary J. Wilson DATE May 24, 2002

TYPED or PRINTED NAME **Mary J. Wilson** REGISTRATION NO. *(if appropriate)* 32,955

☒ Additional inventors are being named on separately numbered sheets attached hereto.

PROVISIONAL APPLICATION FILING ONLY

Burden Hour Statement: This form is estimated to take .2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Office of Assistance Quality and Enhancement Division. Patent and Trademark Office, Washington, DC 20231, and to the Office of Information and Regulatory Affairs, Office of Management and Budget (Project 0651-0037), Washington, DC 20503. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Washington, DC 20231.

631670

PROFILING MOLECULES

The present invention relates to profiling and/or identifying molecules in a sample. It is particularly concerned with profiling chemical or biological molecules contained in an experimental sample using measured data about molecules actually present and known information about candidate molecules that may be present.

In many biological or biochemical situations, such as gene expression analysis and protein mass spectroscopy, it is desired to analyse a sample which contains a mixture of molecular species in order to determine which of a number of possible candidate molecular species are present in that sample. Such analysis may comprise measurement or determination of one or more of properties of molecules in that sample, e.g. absolute number, concentration, fluorescent intensity, etc..

Frequently the information about the possible candidate molecular species is incomplete. For instance there may be other candidate species which are not known about. Alternatively or additionally, the knowledge of the relevant properties of each candidate (such as its DNA sequence, amino acid sequence or atomic composition) may be incomplete.

In order to perform such an assignment of candidate molecules to the contents of the sample, at least one set of measurements needs to be performed for molecules of the sample. For practical chemical reasons, this typically involves measuring the sum total of an activity such as absolute number, concentration or fluorescent intensity, in each of a number of defined fractions. Those fractions may be physical fractions obtained by separating the sample using

physical properties of the molecules (e.g. charge, mass, colour, etc.) or measurement fractions obtained by using different measurement methods or settings. In some preferred embodiments of the present invention, the measurement relates to the amount of nucleic acid fragments arising from cutting transcripts with a given restriction enzyme and which fragments have a given length.

Measurement of each property may not be exact, and there may be an error in the sum total activity. As all such measurements are experimental, they may be subject to errors, which may be errors of fractionation, such as an uncertainty as to which fraction a particular measured species should be placed in or derives from, or of measurement, such as an error as to the precise property or quantity measured. Such errors, if they occur, may complicate resolution of which actual molecules in the sample correspond to which candidate.

In embodiments of the present invention there may be a number of "candidates" which may be contained in a database of known molecular species; a set of "tags" which are the physical properties that are used to define the fractions; and associated "tag quantities" which are the measured quantity of each tag in each fraction. The desire is to create a number of "assignments" between tags and candidates in order to create a profile of the sample under examination and/or identify which candidates are actually present. Some of these assignments may be mutually exclusive, and will be referred to as "mutex assignments" or "links". These words and phrases will be used in this context throughout the specification.

In a broad aspect, the present invention provides a method of assigning tags to candidates. This may be achieved with a high degree of accuracy and a low false positive rate by

minimising the effect of one or more possible sources of error.

In certain embodiments of the present invention an objective goal (assignment) is optimised by linear programming. In other embodiments, which may be preferred, assignment is optimised by mixed integer programming.

In one aspect, the present invention provides a method of providing a profile of molecular species present in a mixture contained in a sample and/or identifying the presence of multiple molecular species in a sample, the method comprising:

generating a dataset comprising information measured or determined for molecules of the sample, including, for one or more fractions out of a plurality of different fractions of molecules wherein each fraction has a different property, a measured or determined sum total of an activity over all molecules that have a particular property, the sum total being assigned to the fraction of molecules with that particular property, and

assigning to molecules for which measured or determined information is present in the dataset a candidate molecular species which may be contained in the sample and for which information is present in a database;

wherein there is uncertainty over at least one of the following:

a) the information for each candidate molecular species in the database,

b) completeness of the database,

c) accuracy of generation of the fractions,

d) accuracy of measurement or determination of the properties,

e) accuracy of measurement or determination of the sum total of the activity;

wherein assigning to molecules for which measured or determined information is present in the dataset a candidate molecular species which may be contained in the sample and for which information is present in the database comprises:

i) generating a set of constraints for a number of candidates based on possible assignments of candidates to measured fractions and resultant limitations on total activity of each candidate potentially present in the sample,

ii) creating an objective function which it is desired to either maximise or minimise,

iii) optimising the objective function with regard to the set of constraints to provide a profile of candidate molecular species actually present in the sample and/or identify candidate molecular species actually present in the sample, and optionally to determining amounts of different candidate molecular species actually present in the sample.

In this context, a profile of molecular species present in a mixture may comprise information on the abundance of each molecular species or on some activity of each molecular species. Such activities may include but are not limited to enzymatic activity or molecular modifications such as protein phosphorylation, or experimentally introduced labels such as fluorochromes or isotope labels.

A method of the invention as claimed may comprise measuring or determining information for molecules of the sample.

A method of the invention as claimed may comprise measuring or determining a sum total of an activity over all molecules that have a particular property (e.g. sum total of abundance of nucleic acid fragments of a particular length).

The objective function may be created in any of various suitable forms, each of which may focus on minimising the impact of a particular form or source of error. For example, an objective function may be used which minimises the part of the sum total quantity in each fraction which is not assigned to candidate species. Alternatively, an objective function may be used which minimises the errors arising from mutually exclusive assignment of candidate species to fractions, as discussed below.

Alternatively a general objective function can be created which contains contributions from each possible source of error, such as those discussed in the previous paragraph, and these contributions can be independently weighted relative to each other using constants.

Preferably some of the constraints take the form of equalities which are constructed using the knowledge that for each tag the quantities of the candidates assigned to the tag together make up the tag-quantity. If the database of candidates is incomplete, then there is a possibility that some of the tag-quantity derives from an unknown candidate. Therefore, an equality can be derived for each tag using a slack variable to represent the amount of the tag-quantity which is not accounted for by the candidates assigned to the tag.

The assignment of candidates to tags may also be ambiguous. For example, as SAGE is based on comparison with sequence databases having an expected error rate of more than 1% per base (or about 10% per 9 base-pair (bp) tag), one may wish to assign each gene to SAGE tags which either match perfectly, or which differ in one or more base positions.

In an example based on fragment analysis (e.g. where the fragments are generated by cutting nucleic acid copied from a sample using restriction enzymes in different fractions, and sorted according to partial sequence information and length/fragment size, optionally with measurement of abundance), ambiguity may occur when measured fragment sizes are not perfectly accurate. If a gene *g* in a sequence database would give rise to a 123 bp fragment, and fragments A and B are observed at 122 and 124 bp respectively with a 1% error in the measurement, then it is impossible to decide on this information alone which fragment should be assigned to the gene.

This ambiguity may be incorporated into the constraints using the fact that the quantity of the candidate is less than or equal to the sum of the tag-quantities associated with all the tags which that candidate could possibly be assigned to. Whilst this is certainly true, it results in a loss of information about the system, namely that a given candidate can in reality only be assigned to one tag, but here it is allowed to contribute a portion to several. In a situation where several tags are closely spaced, or the possible error is comparatively large, this can result in the combining of a significant proportion of the tags in the sample, resulting in insufficient information being available to resolve the possible assignments.

Therefore it is preferable to work from the fact that the assignment of the candidate to each of the tags is normally mutually exclusive (a "mutex assignment") - i.e. it is only possible for each species present to have produced one tag in each experiment on the sample. An embodiment of the present invention proposes a solution to this problem by the introduction of pseudo-Boolean variables (constrained to be

either 0 or 1). These variables are "slack" variables (i.e. they are introduced as a way of expressing an additional constraint) which can be used to express this mutual exclusion in the constraints and thereby create a more powerful model of the true assignment problem.

There are several ways of solving a problem which incorporates Boolean variables, including constraint programming and mixed-integer programming. Mixed integer programming ("MIP") is particularly preferred in embodiments of the present invention, and is significantly more computationally efficient.

Within the framework of mixed integer programming it is also possible to include penalties for assignments by adding terms to the objective function that depend on the selection of assignments by the Boolean variables. Such penalties can be used to express confidence in the assignment, such as in relation to the probability that the assignment should result. For example it may be possible to assign a candidate to several tags, but those that correspond most closely with the known data on the candidate can be made more preferable.

It is normally the case that more than one candidate can be assigned to a tag, but in some cases, it is known that one and only one candidate from a particular group can be assigned to a particular tag. For example, in the case of DNA analysis there may be a sequence ambiguity in the database such that a candidate is known to have either sequence ACGTGC or ACGTGG. In such a case one may introduce both variants as candidates and assign them to tags with the condition that only one of them can be non-zero. This results in a mutual exclusion situation similar to that described above in relation to

assignments, and can be incorporated into the method by similar use of pseudo-Boolean slack variables.

It will be appreciated that the present invention is equally applicable to any situation where it is desired to assign experimental results on a sample to candidate molecular species contained in a database. Table 1 below shows how this method may be applied to three exemplary situations, namely (i) sorting of fragments on the basis of restriction digest, partial sequence information and length ("fragment printing"), (ii) SAGE and (iii) protein mass spectroscopy.

TABLE 1

	Candidate	**Tag**	**Tag-quantity**
(i) Fragment Printing	Transcripts	Fragment subsequence and length	Electrophoresis peak area (abundance)
(ii) SAGE	Transcripts	Fragment subsequence	Subsequence count
(iii) Protein mass spectroscopy	Protein sequences	Peptide mass	Peak Area (abundance of each mass)

Thus, in certain embodiments of the present invention candidate transcripts in a database are assigned to genes actually transcribed in a sample on the basis of determination of partial sequence information ("subsequence") and length of fragments generated in different fractions for the transcribed genes, along with abundance of fragments of determined partial sequence and length. This kind of system is discussed in a great deal more detail further below.

In other embodiments of the present invention, candidate transcripts in a database are assigned to genes actually transcribed in a sample on the basis of partial sequence information ("subsequence") alone - e.g. as in SAGE (Velculescu et al) - along with the abundance of fragments of a determined subsequence measured by counting the number of occurrences of each subsequence in the dataset.

In still further embodiments of the present invention, candidate proteins in a database are assigned to proteins actually present in a sample on the basis of partial structure information obtained by digesting the protein sample into peptides and measuring the mass of each peptide fragment with mass spectroscopy, along with the abundance of each peptide fragment so measured.

Methods for highly parallel gene expression analysis have been available since the early 1990's. Although small-scale methods such as RNase protection and Northern blotting can be scaled up, clearly something else is needed when the goal is to measure the activity of all genes in a sample simultaneously.

Differential display, invented by Liang and Pardee in 1990, was one of the first methods capable of analyzing a large proportion of the expressed genome (Liang et al., 1992, Science 257, 967-971). It is also the first example of an open system, so called because it does not rely on specific primers or probes and is thus not limited to detecting a pre-determined set of genes.

Closed systems rely on gene-specific probes or primers, avoiding the problem of identification. Microarrays, invented by Ed Southern (Southern et al., (1994) J Biotechnol 35, 217-227; Southern et al., (1992) Genomics 13, 1008-1017), introduced a highly parallel closed system, where gene-

specific probes are synthesized or deposited on a surface which is then used to probe an mRNA sample.

Microarrays have been improved in various ways. Brown and colleagues pioneered contact printing of cDNAs (Schena et al., (1995) Science 270, 467-470) while Fodor and colleagues introduced direct oligonucleotide synthesis *in situ* using a mask-based light-directed system (Fodor et al. (1991) Science 251, 767-773), commercialized by Affymetrix as the GeneChip. Since then, oligonucleotide and cDNA microarrays have dominated the market for closed systems. Recently, microarrays carrying oligonucleotides synthesized *in situ* using an ink jet system have become a viable alternative (Hughes et al., (2001) Nat Biotechnol 19, 342-347).

Real-time PCR is a sensitive, accurate and reproducible closed system for gene expression analysis. In real-time PCR target sequences are amplified and the yield of product is monitored during each cycle. The approach permits measuring exactly the PCR amplification efficiency and yield and thereby to calculate the initial concentration of the target. The procedure is expensive, however, and low-throughput, because each gene must be analyzed separately. Hence it is not suitable for global gene expression analysis.

Efforts to find an alternative to differential display have yielded a variety of open systems. Serial analysis of gene expression (SAGE) involves isolating from each transcript a characteristic 9-10 bp tag and then sequencing millions of such tags in order to provide quantitative estimates of gene expression by tag-counting (Velculescu et al., (1995) Science 270, 484-487). While SAGE is an improvement over differential display in that it at least provides a 9-10 bp clue to the identity of each gene measured, such a tag is still not

specific enough to unambiguously identify a gene in most cases. Indeed, statistics published by the SAGEmap project based on over five million tags indicate that only 40% of tags are unique, and that on the order of 10-26% of the tags are due to sequencing errors (Lash et al.,(2000) Genome Res 10, 1051-1060).

A vast improvement over SAGE, Massively Parallel Signature Sequencing (MPSS) generates millions of long (16-20 bp) sequence tags in parallel. MPSS uses an integrated system involving a confocal microscope and microfluidic delivery of a sequence of enzyme mixes to bead-cloned transcripts (Brenner et al.,(2000) Nat Biotechnol 18, 630-634). Because the sequence tags are relatively long, most will be unique even in the human genome. Measuring gene expression then simply reduces to the task of counting tags (disregarding the effect of sequencing errors). However, MPSS remains expensive and low-throughput, averaging between three and eight weeks per sample and instrument.

A number of protocols have been proposed to improve differential display by using adaptor-mediated PCR. Instead of amplifying DNA randomly using low-stringency short PCR primers, these approaches cut the target with restriction enzymes and ligate adaptors containing universal primer templates. In order to display such fragments on a gel or using capillary electrophoresis, the sample must be subdivided into smaller fractions.

One approach, pursued by CuraGen in their GeneCalling method (Shimkets et al.,(1999) Nat Biotechnol 17, 798-803), is to cut with enzymes that recognize quite long sequence motifs, so that only a small number of genes cut with any given enzyme. However, even using 6 bp-recognizing enzymes generates about

800 fragments to be displayed in a window of 1000 bp. The result is mountains, not sharp peaks; as a consequence, low-abundance genes are buried and there has been no way to directly identify the fragments. GeneCalling requires experimental identification of candidates using poison primers.

Another approach, used e.g. in Total Gene Expression Analysis (TOGA) is to subdivide the sample in the amplification step by positioning the primers so that they protrude into the unknown sequence of the fragment (Sutcliffe et al., (2000) Proc Natl Acad Sci U S A 97, 1976-1981). For example, with two primers each protruding two bases into the fragment, four unknown bases can be probed. Using a set of primers representing all possible sequences would yield 256 subreactions. While this approach is valid in principle it suffers from various limitations.

A further approach called DNA Indexing uses a Type IIS enzyme to cut the target. Such enzymes bind a recognition sequence but cut outside it so that a sequence-specific cohesive end is generated. This activity can in principle be used to ligate adaptors consisting of a universal primer template sequence and a cohesive end which is varied between sub-reactions. If a four-base overhang is generated by the enzyme, 256 adaptors would be needed. There have however been problems with finding conditions where adaptor ligation discriminates a four base overhang (Kato (1995) Nucleic Acids Res 23, 3685-3690).

Many gene expression applications involve determining a limited amount of information about each RNA or DNA species present in the sample, including some measure of the quantity (such as a tag count or a fluorescent intensity measured for the fragments in a size-fractionated sample) and some data

related to the identity of each molecule (such as a tag sequence or the position of a band on a gel).

Often, there is a trade-off between obtaining a reliable identification on one hand and cost or throughput on the other. For example, SAGE (serial analysis of gene expression) is a method where a library of short sequence tags, typically 9 bp long, is sequenced. By limiting the tags to 9 bp, multiple tags can be sequenced simultaneoulsy, thus reducing cost to a manageable level. However, 9 bp is normally not enough to uniquely identify a gene corresponding to the tag. In Digital Gene Technology's TOGA, fragments are isolated and displayed on a capillary electrophoresis system in such a way that 8 bp of information is revealed about each fragment. Again, 8 bp is not enough to find a unique candidate in a sequence database.

To solve the problem of identification one can adjust the experimental procedure to reveal more information about each tag or fragment—at the expense of increased cost and decreased throughput.

Another approach is described in UK patent GB-B-2365124 and WO02/08461, where multiple experiments are performed in which each experiment generates a complete set of tag sequences and quantities covering all the DNA molecules in the sample, but the different experiments probe different parts of each molecule so that the tags generated in each experiment are different for each gene. This effectively increases the information available about each gene by generating multiple tags for each gene, not by increasing the size for each tag. An algorithm is then required in order to resolve which tags belong to which gene.

The system is thus a global gene expression analysis method which combines whole-genome scope with direct and robust identification *in silico*. Using DNA indexing, a comprehensive RNA profile is generated in which each gene is represented once, by a single fragment. The indexing procedure reveals a 10 bp sequence tag for each fragment, plus its length. However, such tags can not normally be unambiguously assigned to a gene in a sequence database. A combinatorial identification algorithm, based on three 10 bp tags per gene, is then used to simultaneously quantify and identify virtually all genes.

The approach of GB-B-2365124 and WO02/08461, employing an algorithm that may be called Combinatorial Identification, uses the fact that an assignment of one or more genes to a fragment always provides *some* information about gene activity. For example, if two genes A and B are the only candidates for a single fragment whose abundance was Q, then we can certainly state that A + B = Q. From three complete profiles, we obtain approximately 150,000 such equations but there are only around 50,000 unknowns (the genes).

Combinatorial identification is a procedure that uses length and/or partial sequence information obtained for a set of fragments - where each gene is represented by more than one fragment - to identify in a sequence database those genes (or other sequences) which produced the observed fragments. The key to combinatorial identification is that each gene is probed more than once. This has the consequence that, even though one may find multiple candidate genes for each fragment (as in SAGE), there is collectively enough information to unambiguously identify each gene's contribution to a particular fragment.

The present inventors have realised that improvement can be made, and the improvement is in fact applicable in a variety of systems, as discussed already above and further below.

Brief Description of the Figures

Figure 1a shows the four possible assignments between candidates (transcripts) and tags.

Figure 1b shows the schematic map of the problem of Example I.

Figure 2 shows a portion of a schematic map of an assignment problem.

Figure 3 shows a scatter graph of the experiment described in Example III.

In embodiments of the present invention, transcribed genes in a sample are analysed on the basis of partial sequence information, abundance and length, with different fractions being generated by means of cutting to create fragments using different restriction enzymes and amplification using different adaptors and primers. Details of embodiments of a method of providing a profile for transcribed genes in a sample are found in WO02/08461 and GB patent 2365124.

Tags and tag-quantities produced from the sample may be analysed in accordance with the present invention, employing a database of candidates.

Candidates may first be assigned to the tags. Then any tag-quantity of zero, or of less than a threshold value may be used as evidence that the candidates assigned to it were not

present in the sample at all. These candidates are given an activity of zero, and are removed from all other assignments.

If two or more candidates have exactly the same assignments to tags, they are to be considered as equivalent for the purposes of this analysis. In each such set, all but one of the members are removed from the problem, and a note is made that the members of each such set cannot be distinguished.

The remaining mapping of assignments may include one or more of each of the following:

a) unambiguous assignments - where a single candidate is assigned to a tag;

b) unresolved assignments - where multiple candidates are assigned to a tag;

c) mutex assignments - where a single candidate is linked to multiple tags and it is known that only one of the links is correct - in this case each link may have an associated error measure;

d) multiple assignments - where a candidate is assigned to multiple tags and it is known that all of the assignments are potentially correct (e.g. where the sample has been analysed more than once using different conditions);

e) mutually exclusive candidates - where an uncertainty in the physical properties of a candidate has been modelled by introducing two or more variant candidates with the constraint that only one of them can actually be present in the sample.

The five possibilities a) to e) are shown in a schematic map in Figure 1a.

Therefore the problem can be defined in the following way.

Constants

A tag-quantity Q_p for each tag p.

An error of assignment R_{ag} for each link a in each mutex assignment g. This is an externally generated measure that can capture such things as the difference between observed and expected length of a fragment, the degree of confidence in a tag sequence, etc..

Variables

A decision variable B_{ag} for each link a in each mutex assignment g, which takes the value 0 if the link is not used, and 1 if the link is used. The goal of the procedure is to find a solution where $B_{ag} = 1$ if and only if the link a is the correct assignment.

An expression variable E_t for each candidate t.

An expression variable E_{ag} for each link a in each mutex assignment g, which takes the value 0 if $B_{ag} = 0$ or E_t if B_{ag} is 1.

A slack variable S_p for each tag p, denoting the part of its tag-quantity that is not contributed by the expression E_{ag} of all candidates assigned to it. This slack variable can be used in the objective function, for example to minimise the fraction of tag-quantity not explained by assigned candidates – i.e. minimising S_p/Q_p. S_p must be zero or positive.

Constraints

Each mutex assignment can only have one active link. That is to say that in each mutex assignment at most one decision variable B_{ag} can be non-zero. This is expressed as

$$\sum_a B_{ag} \leq 1$$

and there must be one such constraint for each mutex assignment.

Only the active link contributes a non-zero expression level to the corresponding candidate, as discussed above in relation to the definition of E_{ag}. This is expressed in mixed integer programming by a constraint

$$E_{ag} \leq M \times B_{ag}$$

for each mutex assignment where M is a very large constant (chosen so as to be larger than any possible value of E_{ag}, *i.e. larger than any Q*). There is one such set of constraints for each mutex assignment.

The total contribution of all candidates assigned to a particular tag may not exceed the tag-quantity. This is expressed by stating that the sum of all such contributions to each tag (E_{ap} for all a) plus the remainder S_p must equal the tag-quantity:

$$(\sum_a E_{ap}) + S_p = Q_p$$

There is one such constraint for each tag.

The expression level of each candidate is the same in every assignment it occurs in. In other words, all the E_{ag} for a given candidate must be equal, and can be set equal to the total expression for that candidate E_t. This constraint can be expressed as

$$E_t = \sum_a E_{ag} .$$

This constraint operates with the constraint involving M above to ensure that in each assignment for which a candidate's link is active, its contribution to the tag quantity is its expression E_t, the sum totals of the contributions of all active assignments to a particular tag being constrained to be less than the tag quantity as described in the previous condition.

In the situation where there is a possibility of mutually exclusive candidates, additional variables, and additional constraints involving those variables can be introduced to reflect this. In a similar fashion to the use of pseudo-Boolean decision variables B_{ag} above, an additional decision variable D_i may be introduced for each candidate and similar constraints: $E_i \leq M \times D_i$ and $\Sigma_i D_i \leq 1$ (with one such sum for each set of mutually exclusive candidates) used.

Optional Additional Constraints

In some situations, additional constraints can be useful either to address a particular source of error, or to reflect a known situation. For example, in gene expression analysis, it may be known that one candidate is longer than another, although there is minimal (1 bp) separation between the two. An additional constraint can be used to enforce this proper ordering when assigning these candidates to closely spaced tags. To illustrate more specifically, if candidates C_1 and C_2 are 199 and 200 bp in length, and they are to be assigned to

three tags T_1, T_2 and T_3 at 199, 200 and 201 bp with a known error of ±2 bp, there would normally be nothing to prevent assigning C_1 to T_2 and C_2 to T_1. Such assignments can be explicitly prevented by forcing their decision variables to be mutually exclusive, i.e. by constraining $B_{12} + B_{21} \leq 1$ in this case.

One example of the objective function contains three terms independently weighted using constants K_1, K_2, K_3, wherein the terms are:

a) the negative of the ratio between the tag-quantity which is not assigned to a candidate species S_p and the total tag-quantity Q_p for all tags p - i.e.

$$-\sum_p S_p / Q_p \; ;$$

b) the error of assignment R_{ag} for each active link in each mutex assignment weighted by the expression level and the tag-quantity for all links in all mutex assignments - i.e.

$$\sum_{a,g}((1-R_{ag})\times E_{ag}/Q_g) \; ; \text{ and}$$

c) the negative total error of assignment - i.e.

$$-\sum_{a,g} R_{ag}B_{ag} \; .$$

The resulting compound objective function is therefore given by:

$$-K_1\sum_p S_p/Q_p + K_2\sum_{a,g}((1-R_{ag})\times E_{ag}/Q_g) - K_3\sum_{a,g} R_{ag}B_{ag}$$

By maximising this compound objective function, a best-fit to the sample data can be obtained. The constants K_1, K_2, and K_3

can be used to weight the objective function in favour of one or more of the terms, for example if a particular source of error is known to be small.

It will be appreciated that the same result will be obtained by minimising the negative of the objective function.

The following section provides further description of embodiments of the invention as applied to a tissue sample containing mRNA treated as follows: take a tissue sample, extract mRNA, translate that into cDNA attached to a solid support such as magnetic beads, and then pass them through restriction enzymes and amplification with selected ligation fragments (see e.g. WO02/08461). The restriction enzymes may recognize DNA sequences 5 base pairs long. A more generic name for these recognition sequences is, e.g., a *restriction group*. Ligation fragments may be used that on one side have a four base pair overhang, and on the other have one of the three letters (G,T,C), followed by a poly-A string. That gives 3.4^4 = 768 possibilities. Each of these may be referred to as a *frame*. All the discrete information on fragments, frames and enzymes, may be referred to, e.g., as a *frameset*. In the example set out here, there are 3.768 = 2304 framesets if the transcribed genes are subject to 3 separate restriction digests with different enzymes. Each of these in fact labels one experiment, characterized by which restriction enzyme was used, and which primers were used in the PCR.

Many poly-adenylation sites are not precisely known. It is therefore in any case convenient to define a *head-frame* as the four letters overhang (bases resulting from restriction enzyme digest), and a *sub-frame* as the single letter before the polyadenylation sequence. The *frame* is the combination of the *head-frame* and the *sub-frame*.

Definition 1.1 *a gene (candidate) 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 1.2 *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 1.3 *a peak (tag) is* (ID; enzyme; frame; length; Area), *where* ID *is an identifier provided with the processed experimental data.*

Definition 1.4 *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. It is not supposed to have been listed in the experimental data, but to have identifier* dark. *Since its properties are not observed, it can have any area.*

Definition 1.5 *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).*

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 Figure 2.

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

Definition 1.6 *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* 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 1.7 *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* s[pos - 1].

2 Two sources of uncertainty

The bipartite graph (e.g. that of Figure 2) encodes two sources of uncertainty.

2.1 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 database at all;

2. there are alternative splicings of the same gene, and not all are in the database;

3. there are alternative poly-A sites, and not all are in the database;

4. for a given end, the sequence in the database 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 of 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 $(1 - 1/1500)^{10} \approx 0{:}94$. This rate is of course highly dependent on the accuracy of the genomic sequence, and on the rate of SNPs.

2.2 Fragments-Peaks uncertainty

The second source of uncertainty is that of experimental errors for example in the determination of *length* and *Area* of a peak, and in the comparison with fragments. We can assume that *Area* is reproducible up to a factor about two. True DNA string length is an integer (a whole number of nucleotides), while the observed length is approximately given with one decimal position. (Resolution of the sequencing machine is about 11 data points per base, run-to-run variability is 15% of one base.)

Most of the scatter in *length* of peaks may be 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. This may 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. Thus, these length corrections may be determined and may be included as corrections in a data base.

3 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 Figure 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

$$\Sigma = \{(g,p);\ g \in \{\text{Genes}\};\ p \in \{\text{Peaks}\}\} \qquad (3.1)$$

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

$$\Sigma^t = \{(t,p);\ t \in \{t\text{-Genes}\};\ p \in \{\text{Peaks}\}\} \qquad (3.2)$$

The difference between (3.1) and (3.2) 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:

$$\forall(l=(g,p)) \in \Sigma : start(l) = g\ end(l) = p \qquad (3.3)$$

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

$$enzyme(l) = enzyme(end(l))$$
$$length(l) = length(end(l))$$
$$frame(l) = frame(end(l)) \qquad (3.4)$$

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 3.1 *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 $\sigma(tg, enzyme)$:

$$\sigma : \{(tg, enzyme) \rightarrow (p)\}\ \sigma = p\ (tg,p) \in \Sigma^t \quad (3.5)$$

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 levels. 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.

4 Optimization

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

$$E_{tg} = \text{expression level of t-gene } tg \quad (4.1)$$

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

$$E_p = \sum_{tg:p=\sigma(tg;enzyme(p))} E_{tg} \quad (4.2)$$

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.

We will now make a short detour on target functions; see Press et al., *Numerical Recipes*. Cambridge University Press, 1988, for a longer discussion.

Generally, one would often introduce some function $F(E_p; Area_p; L_p; length_p)$ to minimize. The measured lengths and areas may be considered realizations of random variables. Assume that these random variables have probability distribution functions depending parametrically on σ and the E_{tg} 's. Let this probability be denoted Prob({*Area*}, {*length*}; σ, {*E*}). The *maximum likelihood* method is then to maximize this probability, by varying σ and *E*. It is clearly equivalent to minimizing

$$F = -\log \text{Prob}(\{Area\}, \{length\}; \sigma, \{E\}) \quad (4.3)$$

Every assumed target function can be given an interpretation in terms of an assumed probability density functions, and vice versa.

4.1 All errors independent

Let all measurement errors be independent random variables. Then, with some *f* and *g*,

$$\text{Prob}(Area, length; \sigma, E) = \prod_{\text{t-genes}} \prod_{\text{enzymes}} e^{-f(Lp; lengthp)} \prod_{\text{peaks}} e^{-g(Ep, Areap)} \quad (4.4)$$

and the target function is

$$F = \sum_{\text{t-genes}} \sum_{\text{enzymes}} f(L_p, length_p) + \sum_{\text{peaks}} g(E_p; Area_p) \quad (4.5)$$

One example is a quadratic target function for lengths

$$F_2 = \sum_{\text{t-genes}} \sum_{\text{enzymes}} \alpha_p (L_p - length_p)^2 \quad (4.6)$$

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

$$F_1 = \sum_{\text{t-genes}} \sum_{\text{enzymes}} \alpha_p |L_p - length_p| \quad (4.7)$$

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

A special case to consider is an assignment to a dark peak. One idea would be to estimate the expected distribution of peaks in the range of the dark peak (e.g. around 1200 bp, for example) and then calculate the expected distance from an average such peak as above using some distribution. That can be done by finding the known frequency of fragment lengths in the database across the entire range of lengths (i.e. beyond 1000 bp), fitting this to a curve (1/x) and calibrating this using the observed frequency of peaks in the observable region (i.e. below 1000 bp). From this calibrated curve one can obtain the local expected frequency of peaks across all lengths. This implicitly gives the expected distance to the nearest peak and thus the probability of the length offsets. Similarly one could give an *a priori* probability of an assignment to a zero peak, by estimating the number and

distribution of sub-threshold peaks and calculating the distance from an average such peak. Use the calibrated curve from above, which essentially describes the local frequency of above-background peaks. Given an estimate of the proportion of genes which ought to be above background, we can then determine the local frequency of below-background peaks. This proportion can perhaps be determined experimentally.

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

$$F_{ln} = \sum_{\text{peaks}} s_p \ (\log E_p/Area_p)^2 \qquad (4.8)$$

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.

4.2 Errors from one gene correlated

One can also employ 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 p1; p2; p3 for a given gene, one would then have

$$F_{corr} = \sum_{\text{t-genes}} h(L_{p1}, length_{p1} \ldots, E_{p1}, Area_{p1} \ldots) \qquad (4.9)$$

This formulation is particularly straightforward for quadratic target functions.

Let $\rightarrow\Delta l$ be the vector of errors in length, with different enzymes, and M_{tg} be a square matrix, then

$$F_{2,corr} = \sum_{\text{t-genes}} c_{tg}(\rightarrow\Delta l.\ M_{tg}\rightarrow \Delta l) \qquad (4.10)$$

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

$$((length_{pi} - L_{pi})(length_{pj} - L_{pj})) = (M_{tg}^{-1})_{ij} \quad (4.11)$$

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

5 A constraint model for a gene matching problem

This section describes a constraint model of the problem defined in the preceding sections. There is a certain overlap between the terminology developed here and that used in section 3. Where concepts introduced here are identical to or related to those of section 3 this has been noted in the text.

5.1 Fragment Lengths

Definition 5.1 *Represent the expected FRAGMENT LENGTH (in discrete base pairs) of all known t-genes as a n x m matrix where each column represents the expected lengths of the final fragment of the i:th t-gene from the last occurrence of the recognition sequence to the end of the t-gene for each of the* m RESTRICTION GROUPS.

Let a value of 0 for L_{ij} in the matrix represent that the recognition sequence of restriction group j does not, up to the reliable scope of the experimental equipment, occur in t-gene i.

Note 1 *If a gene has a several known polyadenylation sites we represent it as separate t-genes with different expected fragment lengths. Similarly, if there is a range of possible adenylation sites for a known gene and regions within this range which result in completely different fragments or in widely differing fragment lengths in at least one of the restriction groups, we represent each such region as a separate "t-gene" with fragment lengths corresponding to the most likely polyadenylation site within the region.*

5.2 Frames and sub-frames

5.2.1 Head-frame and sub-frame mappings

The expression of each t-gene will occur either in a particular head-frame or not at all in each of the restriction groups and since we know the sequence of the (known) t-genes we can define a MAPPING from a t-gene *i* and a restriction group *j* to the initial sequence of the final fragment for that restriction group.

Definition 5.2 *Let Φ be a function from a t-gene i and an restriction group j to the initial sequence of the fragment immediately following the last occurrence of the recognition sequence in the t-gene.*

Note 2 *This means that for any given t-gene i and restriction group j, Φ (I,j) represents the head-frame in which we should expect to find its expression.*

Since the polyadenylation site of a t-gene *i* may not be exactly known, the expected fragment length given by L_{ij} for restriction group *j* may be off by one or more bases. We cannot therefore predict with certainty the sub-frame in which the expression of a particular t-gene can be observed. We can however for each t-gene and restriction group give the sub-frame for any given offset from the expected fragment length.

Definition 5.3 *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 subframe* ς(i,o) *(of head-frame* $\Phi(i,j)$*).*

Note 3 *Because of the representation choices made in the note to definition 5.1 we can expect the actual fragment length to be close to the expected length given by* L. *The function* ς *therefore does not have to encode the complete sequence for all known* t-genes *but only the sequence in a region around the fragment length for each* t-gene.

5.3 Experiments

Definition 5.4 *Represent the* RESULT OF AN EXPERIMENT *as a finite list of peak parameters*

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

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

Notation 5.5 *Let* τ *represent the basepair resolution of the experimental equipment and let* $\pi_1 = \pi(1)$ *to denote the observed position of an observed peak given in discrete units*

corresponding to one T : th *of a base pair and* $\alpha_1 = \alpha(1)$ *the observed area in arbitrary units and let furthermore* $\gamma_1 = \gamma(1)$ *denote the restriction group,* $\Phi_1 = \Phi(1)$ *the head-frame and* $\varsigma_1 = \varsigma(1)$ *sub-frame of each peak* P1.

(Note that π and γ correspond to the *length* and *enzyme* function of section 3 and that the *frame* function of that section is here split into the two functions Φ and ς.)

Note 4 *Since* P *is sorted by peak position, we know that the following condition holds*

$$\forall kl(k<l \Rightarrow \pi(k) \leq \pi(l))$$

5.4 Matchings

Since we know that several t-genes may contribute to a particular peak and conjecture that a t-gene can contribute to at most one fragment length in each restriction group we 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 5.6 *Represent a* MATCHING *as 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* $C_{ij} = 1$ *for some* $0 < 1 \leq p$ *otherwise.*

(Note that the notion of matching defined here is very close to that of section 3. We have here one variable C_{ij} for each t-gene i and restriction group j. Assigning a value 1 to one of these variables corresponds exactly to letting the value of the expression $\sigma(i,j)$ be 1.)

5.4.1 Necessary conditions on matchings

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

Condition 5.7

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

i.e. 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:

Condition 5.8

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

and

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

Let each variable O_i have initial domain constrained by

Condition 5.9

$$\forall(i \leq n)\ (\lambda_i \leq O_i \leq v_i)$$

where λ_i and v_i impose hard limits on the offset from the expected polyadenylation site for a given t-gene i.

Note 5 *We let λ_i and v_i be parameters that depend on the probability of finding the polyadenylation site for t-gene i at an offset O_i from the one used to compute the expected*

fragment lengths L_{ij} in each of the restriction groups. Note that with the definition of t-gene used we can expect the value of O_i to be quite close to zero. A first guess for λ_i would be between -5 and 0 and for v_i between 0 and 15 depending on the confidence put in the expected fragment length L_{ij}. Asymmetry of the parameters can be used to encode the case where we know e.g. the first likely polyadenylation site but are uncertain if there are more.

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*.

Condition 5.10

$$\forall(i \leq n)\ (\exists_j\ (C_{ij} = 0 \varpi L_{ij} \neq 0) \Rightarrow (O_i = 0 \varpi \forall(j \leq m)\ (C_{ij} = 0)))$$

5.5 Matching Errors

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 that we will use to enforce further necessary conditions on the matching and a matching penalty.

Definition 5.11 *Let the* MISMATCH ERROR Me_{ij} *of a particular* t-gene i *in a restriction group* j *be defined as*

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

Note 6 *This error captures for each* t-gene *and restriction group the (positive or negative) distance from the expected position* $\tau L_{ij} + \tau O_i$ *from the position* $\pi(C_{ij})$ *of peak* C_{ij} *for any given offset* O_i *in units corresponding to one* τ : th *of a base pair.*

5.5.1 Necessary conditions on matching errors

Since Me_{ij} encodes the absolute offset from the expected fragment length we can use it to constrain the matching variables as follows:

Condition 5.12 *Let each* C_{ij} *have an initial domain* 0, . . . , p *and constrain the* C_{ij} *and* O_i *variables by enforcing for each* t-gene i *and restriction group* j *the following condition*

$$\lambda_{ij} \leq Me_{ij} \leq v_{ij}$$

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

Note 7 *Note that since we chose to let the peak indexes denote peak position order, condition 5.12 can be efficiently enforced and checked using the* element/3 *constraint of a constraint programming system.*

We let λ_{ij} and v_{ij} *depend on the sequence dependent probability of detecting a fragment of* t-gene i *at position* $\pi(C_{ij})$ *in restriction group* j. *A first guess is that typical values for* λ_{ij} *should be between* -1τ *and* -15τ *depending on fragment length (and possibly actual sequence) and for* v_{ij} *between 0 and* 2τ *depending (probably only) on fragment length. The asymmetry of the parameters is motivated by the observation that sequence*

dependent errors usually manifests as apparently shorter fragments lengths.

Correlations of errors between the restriction groups for a particular t-gene indicate sequence dependent systematic offsets from the expected fragment length. Such a correlation should depend statistically on relative fragment lengths. Because of this, the closer fragments are in length, the less discrepancy in error we should tolerate. Motivated by this observations will enforce limits on the discrepancies between the errors in the restriction groups for any given t-gene *i*.

Condition 5.13 *For each* t-gene i *and any two restriction groups* j *and* k *such that* $C_{ij} \neq 0$ *and* $Cik \neq 0$ *enforce a condition of the following form*

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

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

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*:

Condition 5.14

$$\sum_{j \leq mm Cij \neq 0} |Me_{ij}| < \mu_i$$

where μ_i *represents a maximum acceptable total error* over the m *restriction groups for* t-*gene i.*

5.6 Matching penalty

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

$$M_{pi} = \sum_{j \leq mm Cij \neq 0} \mu_{ij}(Me_{ij})_i + \sum_{j=1}^{m-1} \sum_{k=j}^{m} (\delta_{ijk}(Me_{ij} - Me_{ik})) + \kappa_i |O_i|$$

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* (j,k) *of restriction groups and* κ_i *is a constant parameter encoding the relative penalty given to offsets in polyadenylation site for* t-gene *i*.

(The μ_{ij} should inspect both the sign and size of $Me_{ij} - O_i$ and probably depend on the expected fragment length in each enzyme group.

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*.

κ_i should probably depend on the confidence we put the adenylation site used to compute the expected fragment lengths for gene *i*.)

Note 8 *Note that this penalty does not take into account any quantitative measures at all. In principle we can make an assignment of* 0 *to all* C_{ij} *variables which 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.*

5.7 Quantitative expression

Definition 5.16 *Represent the quantitative level of expression of the* i*:th* t*-gene with a variable* E_i *in the same units as that used to express the area* α_l *of each peak* P_l.

(The E_i variables correspond exactly to the E_{tg} variables of section 4.)

Make the following observations:

1. The sum of contributions to a particular observed peak may not (up to the sum 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.

5.7.1 Necessary conditions on quantitative expressions

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

Condition 5.17 *Enforce for each observed peak* P_l *and any given matching the following condition on the expression variables:*

$$\forall (1 \leq p) \quad ((\sum_{\{i \mid C_{i\gamma(1)}=1\}} E_i) \leq \zeta_1 \alpha(1))$$

where ζ_1 *is a parameter representing a maximum acceptable* AREA UNDERESTIMATE *of the area of the peak 1. We let* ζ_1 *depend on the position* π_1 *of peak* P_1.

Note 9 *A value of 2 for* ζ *allows a factor two of area underestimate of a peak in the experiment which appears to be a realistic first estimate. This formulation does not take care of systematic errors such as failed capillaries or ligation errors giving rise to "bleeding" between sub-frames etc. Such errors may be compensated for in input data.*

(A failed capillary could be represented as having peaks of sufficient size in all discrete positions.)

5.7.2 Quantitative errors

To be able to penalize both area underestimate 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 5.18 *Let the* ABSOLUTE QUANTITATIVE ERROR Qe_1 *for each observed peak* 1 *and any given matching be defined by*

$$Ae_1 = \alpha(1) - (\sum_{\{i \mid C_{i\gamma(1)}=1\}} E_i)$$

Note 10 *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 underestimate the absolute error will be negative.*

5.7.3 Quantitative penalties

Definition 5.19 *Let the area underestimate penalty Ap_l for a particular peak P_l be defined by the following expression*

$$Ap_l = \begin{cases} Ae_l & \text{if } Ae_l \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Note 11 *This penalty is defined as the positive contribution of the absolute quantitative error.*

Definition 5.20 *Let the area underestimate penalty Up_l for a particular peak P_l be defined by the following expression*

$$Up_l = \begin{cases} -Ae_l/\alpha_l & \text{if } Ae_l < 0 \\ 0 & \text{otherwise} \end{cases}$$

5.7.4 Total penalty

We will now define a cost function to replace the one suggested in definition 5.15 above with one based on the quantitative expression error as follows:

Definition 5.21 *Let the* TOTAL PENALTY *for a given matching and a given assignment of the expression variables E_i be defined by the following expression:*

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

where $\acute{\omega}$ and ω are suitable weights expressing the relative contribution to the penalty from area underestimate and unexplained area of observed peaks respectively

5.8 Optimal matching

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 definition 5.21 subject to the conditions in sections 5.4.1, 5.5.1 and 5.7.1.

5.9 Notes on search

5.9.1 Clustering

Note that for purposes of search we can decompose the problem into independent subproblems as follows.

Definition 5.22 *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 of sections 5.4.1 and 5.5.1.*

Note 12 *The size of the set of candidate peaks depend heavily on the parameters used in the necessary conditions for matches and match errors.*

Definition 5.23 *Let furthermore the* CANDIDATE t-GENES *of a peak* P_1 *be all the* t-genes *whose candidate peaks include* P_1.

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

5.9.2 Assignment utility

The utility of an assignment of all variables associated with a particular t-gene can be assessed by considering its relative contribution to the areas of peaks assigned to it.

This could be used to formulate heuristics for search of optimal assignments.

Definition 5.25 *Let the relative contribution of a t-gene i to the peaks it is assigned to be defined by the following expression*

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

6 An mixed-integer programming approach

In this section, we describe a mixed-integer programming (MIP) model expressing the optimization problem. The model is closely related to the bipartite graph view; see Figure 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 $sub\text{-}frame(ug) \neq sub\text{-}frame(p)$.

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, which we take as a weighted sum of the peak area error plus the peak length error. 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. To

estimate the peak length error, we assume a function $\rho(ug,p)$ which gives the penalty of $(ug,p) \in \sigma$.

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. The model decomposes into independent subproblems, one per connected component of Σ^u.

6.1 Variables

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

- A decision variable $B_{ug,p}$ for each link $(ug,p) \in \Sigma^u$ such that $B_{ug,p} = 1$ iff $(ug,p) \in \sigma$.

- An expression variable E_{ug} for each u-gene.

- An expression variable $E_{ug,p}$ for each decision variable, the value of which is either 0 or equal to E_{ug}. This is an artifact of the requirement that the MIP model be linear and finitary.

6.2 Objective function

We formulate the model as a maximization problem. The objective is to maximize the function:

$$\sum_{ug} E_{ug} - \sum_{ug,p} \rho(ug,p) \cdot B_{ug,p} \qquad (6.1)$$

6.3 Constraints

Valid assignments. Every u-gene can contribute to at most one peak per enzyme.

Here, M is a very large constant (chosen so as to be larger than any possible value of E_{ug}).

$$\forall ug, z : \sum_{p:enzyme(p)=z} B_{ug,p} \leq 1 \qquad (6.2)$$

$$\forall ug,p : E_{ug,p} \leq M.\ B_{ug,p} \qquad (6.3)$$

No peak overcoverage.

$$\forall p : \sum_{ug} E_{ug,p} \leq Area(p) \qquad (6.4)$$

Consistent expression levels. The expression level of a u-gene is bounded above by the expression level admitted by each enzyme.

$$\forall ug,\ z : E_{ug} \leq \sum_{p:enzyme(p)=z} E_{ug,p} \qquad (6.5)$$

Redundant constraints. Introducing a decision variable B_{ug} for each u-gene, the following implied constraints can be added, subsuming constraint 6.2 (although not necessarily with any computational benefit).

$$\forall ug,\ z : \sum_{p:enzyme(p)=z} B_{ug,p} = B_{ug} \qquad (6.6)$$

$$\forall_{ug} : E_{ug} \leq M\ .\ B_{ug} \qquad (6.7)$$

6.4 Underdetermined problems

Ignoring peak length errors, a subproblem may easily be underdetermined. That is, the subproblem may 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 propose to identify each such group ω of u-genes and to replace its expression variables by a single expression variable for the whole group. The idea is to reflect the ambiguous evidence in the solutions reported. We now present one approach to handling underdetermined problems:

1. Let two u-genes belong to the same group iff they have the same sets of incident links.

2. Solve a modified maximization problem, yielding a total expression E_ω as well as the group's contribution $E_{\omega,p}$ to each peak p.

3. Determine lower and upper bounds, consistent with step 2, of the expression levels of the u-genes and genes that constitute each group.

Let $\rho(\omega,p)$ be defined as $\min_{ug\in\omega} \rho(ug, p)$. We modify the original optimization problem as follows:

- Throughout, we replace ug by ω.

- We replace constraint 6.2 by:

$$\forall \omega, z : \sum_{p:\text{enzyme}(p)=z} B_{\omega,p} \leq |\omega| \qquad (6.8)$$

To determine lower and upper bounds of expression of individual u-genes and genes that constitute a group, we must again formulate and solve an optimization problem.

Objective function. To obtain expression bounds of a u-gene ug, minimize or maximize E_{ug}. note that, in the absence of

extra combinatorial constraints, the expression levels will have the same bounds for all u-genes. To obtain lower and upper bounds for a gene g, minimize or maximize

$$\sum_{ID(ug)=ID(g)} E_{ug}.$$

Variables. $B_{ug,p}$, $E_{ug,p}$, E_{ug} as in the original problem.

Constraints. Constraints 6.2 and 6.3 as in the original problem. Constraint 6.5 is tightened to an equality. Finally, constraint 6.4 is tightened to:

$$\forall p : \sum_{ug} E_{ug,p} = E_{\omega,p} \qquad (6.9)$$

$$\sum_{ug} E_{ug} = E_{\omega} \qquad (6.10)$$

6.5 Final remarks

Further constraints on valid assignments can be readily expressed as constraints on the $B_{ug,p}$ variables.
Many MIP solvers handle so called SOS Type 1 constraints such as constraint (6.2) specially. CPLEX (Ilog, Mountain View, CA, US), 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 $B_{ug,p}$ could e.g. be chosen as $\rho(ug,p)^{-1}$.

Ignoring peak length errors, the problem can be thought of as the following constrained max-flow problem:

- introduce a node $V_{ug,z}$ for each u-gene in the graph and each enzyme, and a link from each node to its neighbor peaks;

- introduce a source node connected to all V nodes;

- introduce a sink node connected to all peaks;

- each node $V_{ug,z}$ can have non-zero flow on at most one outgoing link;

- for each u-gene *ug*, the flow from the source node to each $V_{ug,z}$ node must be equal.

7 Local search

This section describes a local search approach, and an heuristics by analogy with statistical mechanics of disordered systems.

The idea is to treat length and area errors separately. Here is an itemized description of a suggested procedure, except the linear programming part, which substantially overlaps with the mixed integer approach in section 6.

- Start with the assignment with smallest length error.

Comment: In fact, such an assignment is perhaps not substantially easier to find than the overall optimal solution. Alternatively one could start with just some feasible assignment.

- Given that multiple t-genes can be assigned to the same peak, we can form connected clusters of genes which collectively are assigned to a set of peaks to which no other genes are assigned. Each assignment thus defines a set of clusters. To each cluster we can assign an error of assignment E_a, given by the average assignment probability score of the genes in the cluster, suitably normalized.

- Each cluster can be independently solved to yield the gene activities. Least-squares optimization or other error functions may be used. Given a (possibly partial) solution, one may calculate the error of quantification E_q, given by the difference between observed and calculated peak heights.

- Together, the two sources of error can be described as $E = a \cdot E_a + b \cdot E_q$, where a and b determine the relative emphasis on length and area errors.

- A given solution can be improved step-wise. Pick a cluster (for example, the one with the largest error E) and redo the peak assignment for one or more t-genes in the cluster (for example, those corresponding to the peaks with the largest errors). For example, pick one t-gene assigned to the peak with the largest error in the cluster with the largest error. Re-assign it by picking the second-most-probable link from its set of possible links. When running out of new links, pick another gene assigned to the same peak, then pick the second-most erroneous peak etc.

- Changing the links in an assignment of t-genes in a cluster will change the connectivity of that cluster. It may become divided into two or more separate clusters, and it may become linked with a previously unrelated cluster. However, the change will be local and most other clusters will be unaffected.

- Re-solving only the changed clusters will provide new error measures for these newly formed clusters.

- This procedure can be iterated until convergence.

7.1 Analogy with statistical mechanics

The above method of local search assumes that a change of assignment is accepted, if the combined score decreases. This might lead to getting stuck in a local minimum. In statistical mechanics, such problems are often treated with *simulated annealing*, which can just be considered a particular search heuristics. We will provide a short introduction to simulated annealing, and a suggestion how it can be used in the problem.

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 section 6. The state variables are the link spins. As auxiliary variables, we have the gene expression levels of (4.1). A set of link spins is identified with an assignment σ according to

$$\forall l = (tg,p) \in \Sigma^t \; ((s_l = 1) \equiv (l \in \sigma)) \qquad (7.1)$$

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 section 4 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 $e^{-E(\{s\})/T}$, where 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 (4.3). We introduce the concepts *state sum* and *free energy*

$$Z(T) = \sum e^{-E/T} \qquad F(T) = -T \log Z(T) \qquad (7.2)$$

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

$$\text{Min } E = \text{Lim}_{T \to 0} F(T) \qquad (7.3)$$

Equations (7.2) and (7.3) 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 changes if they lower energy
- accept the change with probability $e^{-\Delta E/T}$ if the energy change ΔE is positive
- generate enough changes such that the average free energy at this temperature has stabilized
- lower temperature and loop

In relation to implementing a simulated annealing procedure, consideration needs to be given to 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

(Press et al., *Numerical Recipes*. Cambridge University Press, 1988).

Further aspects and embodiments of the present invention will be apparent to those skilled in the art. All documents mentioned herein are incorporated by reference.

Example I

The following example shows a sample input file for the set of three tags shown in schematic map form in Figure 1b, along with the resulting mixed-integer programming (MIP) problem and the result.

i) Input data in XML format:

```
<cluster objective-value="0.0">
 <candidates>
  <candidate id="1" error="0.0" active="false" activity="0"/>
  <candidate id="2" error="0.0" active="false" activity="0"/>
  <candidate id="3" error="0.0" active="false" activity="0"/>
 </candidates>
 <tags>
  <tag id="1" quantity="10.0" remainder="0"/>
  <tag id="2" quantity="20.0" remainder="0"/>
  <tag id="3" quantity="30.0" remainder="0"/>
 </tags>
 <assignments>
  <assignment candidate="1" tag="1" error="0.0" active="false" activity="0"/>
  <assignment candidate="2" tag="1" error="0.0" active="false" activity="0"/>
  <assignment candidate="2" tag="2" error="0.0" active="false" activity="0"/>
  <assignment candidate="3" tag="2" error="0.5" active="false" activity="0"/>
  <assignment candidate="3" tag="3" error="0.2" active="false" activity="0"/>
 </assignments>
 <assignment-mutexes>
  <assignment-mutex>
```

```
    <assignment candidate="3" tag="2"/>
    <assignment candidate="3" tag="3"/>
   </assignment-mutex>
  </assignment-mutexes>
</cluster>
```

ii) The MIP problem derived from the above input file:

```
maximize
        0.1E_1_1 + 0.1E_2_1 + 0.05E_2_2 + 0.025E_3_2-0.5B_3_2 +
0.0266666666667E_3_3-0.2B_3_3-0.1S_1-0.05S_2-0.0333333333333S_3

subject to
        E_1_1-1000000.0B_1_1 <= 0
        E_2_1-1000000.0B_2_1 <= 0
        E_2_2-1000000.0B_2_2 <= 0
        E_3_2-1000000.0B_3_2 <= 0
        E_3_3-1000000.0B_3_3 <= 0
        E_3-E_3_2-E_3_3 = 0
        E_1_1-E_1 = 0
        E_2_1-E_2 = 0
        E_2_2-E_2 = 0
        S_1 + E_1_1 + E_2_1 = 10.0
        S_2 + E_2_2 + E_3_2 = 20.0
        S_3 + E_3_3 = 30.0

bounds

binaries
        B_1_1
        B_2_1
        B_2_2
        B_3_2
        B_3_3
```

```
SOS
    s1:: B_3_2 B_3_3

end
```

iii) Solution:

E_3	30.000000
E_2	10.000000
E_2_1	10.000000
E_2_2	10.000000
E_3_3	30.000000
B_1_1	1
B_2_1	1
B_2_2	1
B_3_3	1
S_2	10.000000

(All other variables are zero)

(This is the actual data input and output from the commercial MIP solver CPLEX (Ilog, Mountain View, CA, US).)

Example II

The MIP approach has also been tested by generating a simulated dataset based on based on a mouse gene database. The database contained approximately 18,000 known genes, derived from the non-redundant set of genes that can be obtained from Refseq, Unigene, Riken and Genbank.

We generated a simulated dataset containing the known genes and an additional 21,000 genes for a total of 39,000 genes. Unique transcripts were generated from these genes by applying

two polyadenylation variants per gene. The dataset was used to generate simulated raw data *in silico*, with fragment size errors of 3 bp + 1.5%. The algorithm described above was then used to identify and quantify the 18,000 known genes from this simulated raw data. Penalties (R-variables) were assigned based on the distance from observed to expected fragment length.

The algorithm was capable of correctly identifying 99% of the ~18,000 genes designated as "known", with a false-positive rate of 8%. The result was also quantitatively correct: in 96% of the cases, the expression level of the gene was within 2% of the correct value. The results of a number of additional simulations with different parameter settings are shown in Table 2.

Example III

We have also performed experiments on real samples. An experiment performed essentially as described in WO02/08461 was submitted for analysis using the present invention.

7.5 µg mouse muscle mRNA was analyzed using three different Type IIS enzymes to generate three complete fragment profiles, each consisting of approximately 30,000 fragments in a total of 768 subreactions. Due to the experimental protocol, each fragment revealed 10 bp of sequence plus its length.

Database construction

A sequence database was constructed from several sources, including annotated GeneBank sequences, RIKEN full length cDNAs, Unigene and RefSeq as well as some internally generated 3'-end cDNA sequences.

Sequences were aligned to each other, and when possible, ambiguities were corrected. Sequences that had a closed overlap were joined into a longer contig.

PolyA sites were assigned using annotated data, polyA tails, alignment of 3' ends, the occurrence of polyA signals and additional accumulated knowledge.

PolyA sites were scored according to how strong evidence that support them, and the scores were used in optimizations during the combinatorial identification.

When a polyA site was uniquely determined as an effect of the algorithm analysing sample data, this polyA site was added to the knowledge base, hence improving the quality of the gene database.

The database contained ~48 000 transcripts. Only the ~18 000 transcripts which had good evidence for an open reading frame coding for a protein were used in the following analysis.

Peak linking

Peak linking is the process of building a model that represents all possible assignments of genes to peaks. The model can then be optimized using various methods to find an assignment (and thus a gene expression profile) that minimizes some measure of the error.

The algorithm starts with a file containing all peaks, each obtained in a particular sub-reaction which determines the frame ("subsequence") and restriction enzyme used. So for each

peak, the size (in base pairs) of the corresponding fragment is known to within ~2 bp, together with a 9 bp sequence tag.

In a first pass, the database was scanned and gene candidates were assigned to each peak. For each such assignment, there can be additional constraints. For example, a gene may be a candidate for one or the other of two nearby peaks, but not both at the same time (since it has a definite length). Such constraints were compiled as described above and added to the model.

Optimization

The model was then submitted to a combinatorial solver (Ilog CPLEX) which optimizes an error measure. The objective was to minimize the assignment error (i.e. the difference between observed and expected fragment sizes) while maximizing explanatory power (i.e. the extent to which the pattern of peaks is 'explained' by the calculated gene expression profile). The constraints and objective function were as described above.

After optimization, there were a number of left-over peaks, to which no gene was assigned. Such peaks largely represent novel genes and gene variants, although some of them may also represent experimental background (PCR artefacts, misligations etc.).

After optimization, all gene candidates that entered into the optimization were assigned values corresponding to their abundance at the optimum of the objective function. Based on the simulations above, about 17 800 correct (to within 2%) estimates and 1440 false positives were expected. Figure 3 is a scatterplot of a replicated experiment, where 511

transcripts that were detected in both experiments are plotted against each other, illustrating the reproducibility of the method. The variance observed was not greater than that observed for the raw data (electrophoretic peaks), providing indication that the method of this invention adds insignificantly to the experimental error while affording a direct and robust identification and quantification of molecules present in a sample.

TABLE 2

Simulated genes	Detected genes	Sensitivity	Specificity
3391	2950	87%	97%
3464	3001	87%	88%
3462	3020	87%	94%
3488	3055	88%	83%
8584	7626	89%	97%
8452	7452	88%	89%
8510	7505	88%	95%
8540	7563	89%	86%
3391	2942	87%	98%
3464	2992	86%	89%
3462	3030	88%	94%
3488	3070	88%	82%
8584	7582	88%	98%
8452	7461	88%	89%
8510	7584	89%	95%
8540	7633	89%	86%
3391	2962	87%	97%
3464	3011	87%	87%
3462	3056	88%	88%
3488	3086	88%	75%
8584	7688	90%	97%
8452	7729	91%	86%
8510	7760	91%	90%
8540	7690	90%	82%
3391	2952	87%	98%
3464	3004	87%	88%
3462	3043	88%	92%
3488	3076	88%	81%
8584	7653	89%	98%
8452	7518	89%	88%
8510	7620	90%	94%
8540	7659	90%	84%

Table 2 shows results of a large number of simulations involving a set of genes ('Simulated genes') that were treated *in silico* to generate three fragments each in accordance with WO02/08461. Each fragment was quantified and sized and a 9 bp tag was obtained. Noise was added to both the quantity and the fragment size. The resulting dataset was then analysed by means of algorithm as described in Example II, identifying and quantifying the genes in the sample ('Detected genes'). The detection ratio ('Sensitivity') and the inverse of the false-positive ratio ('Specificity') are shown.

CLAIMS:

1. A method of providing a profile of molecular species present in a mixture contained in a sample and/or identifying the presence of multiple molecular species in a sample, the method comprising:

generating a dataset comprising information measured or determined for molecules of the sample, including, for one or more fractions out of a plurality of different fractions of molecules wherein each fraction has a different property, a measured or determined sum total of an activity over all molecules that have a particular property, the sum total being assigned to the fraction of molecules with that particular property, and

assigning to molecules for which measured or determined information is present in the dataset a candidate molecular species which may be contained in the sample and for which information is present in a database;

wherein there is uncertainty over at least one of the following:

a) the information for each candidate molecular species in the database,

b) completeness of the database,

c) accuracy of generation of the fractions,

d) accuracy of measurement or determination of the properties,

e) accuracy of measurement or determination of the sum total of the activity;

wherein assigning to molecules for which measured or determined information is present in the dataset a candidate molecular species which may be contained in the sample and for which information is present in the database comprises:

i) generating a set of constraints for a number of candidates based on possible assignments of candidates to

measured fractions and resultant limitations on total activity of each candidate potentially present in the sample,

ii) creating an objective function which it is desired to either maximise or minimise,

iii) optimising the objective function with regard to the set of constraints to provide a profile of candidate molecular species actually present in the sample and/or identify candidate molecular species actually present in the sample, and optionally to determining amounts of different candidate molecular species actually present in the sample.

2. A method according to claim 1 comprising optimising the objective function by linear programming.

3. A method according to claim 1 comprising optimising the objective function by mixed integer programming.

4. A method according to claim 1 wherein the objective function includes a term dependent on an amount of measured total activity in each fraction which is not accounted for by the candidate or candidates assigned to molecules in that fraction.

5. A method according to claim 4 wherein the objective function includes a term proportional to

$$-\sum_{p} S_p / Q_p$$

where S_p is activity not accounted for by the candidate or candidates assigned to molecules in a fraction p and Q_p is the sum total activity in a fraction p.

6. A method according to claim 1 wherein at least one of said possible assignments of candidates to molecules is

mutually exclusive with at least one other of said possible assignments.

7. A method according to claim 6 wherein the or each mutually exclusive assignment is represented in said set of constraints by use of Boolean or pseudo-Boolean variables.

8. A method according to claim 6 wherein each possible assignment in the or each mutually exclusive assignment is accorded an error value dependent on likelihood that that possible assignment is correct.

9. A method according to claim 8 wherein the objective function includes a term dependent on said error value or values.

10. A method according to claim 9 wherein the objective function includes a term proportional to

$$\sum_{a,g}((1-R_{ag})/Q_g)\times E_{ag}$$

where R_{ag} is error of assignment for each link a in each mutually exclusive assignment g, Q_g is sum total activity accorded to the assignment g and E_{ag} is a non-negative variable taking the value of the amount of the candidate molecular species if link a is correct and 0 if link a is incorrect.

11. A method according to claim 9 wherein the objective function includes a term proportional to

$$-\sum_{a,g}R_{ag}B_{ag}$$

where R_{ag} is error of assignment for each link a in each mutually exclusive assignment g and B_{ag} is a pseudo-Boolean variable taking value 1 if link a is correct and 0 if link a is incorrect.

12. A method according to claim 1 wherein the objective function is a compound function comprising a plurality of functions each of which is dependent on at least one of sources of error which it is desired to minimise, each function within the compound function being weighted by independent constants.

13. A method according to claim 12 wherein the compound objective function takes the form

$$-K_1\sum_p S_p/Q_p + K_2\sum_{a,g}((1-R_{ag})/Q_g)\times E_{ag} - K_3\sum_{a,g} R_{ag}B_{ag}$$

where K_1, K_2 and K_3 are weighting constants, S_p is activity not accounted for by the candidate or candidates assigned to molecules in a fraction p, Q_p is sum total activity in a fraction p, R_{ag} is error of assignment for each link a in each mutually exclusive assignment g, Q_g is sum total activity accorded to assignment g, E_{ag} is a non-negative variable taking the value of the amount of the candidate molecular species if the link a is correct and 0 if the link a is incorrect and B_{ag} is a pseudo-Boolean variable taking the value 1 if the link a is correct and 0 if the link a is incorrect.

14. A method according to claim 1 wherein at least some of said set of constraints are generated by using the sum total of the measured activity in each fraction to constrain the activity of the candidates potentially assigned to that fraction.

15. A method according to claim 14 wherein said constraints generated using the sum total of the measured activity in each fraction take the form

$$(\sum_a E_{ap}) + S_p = Q_p$$

where Q_p is the measured sum total activity in fraction p, E_{ap} is a non-negative variable taking the value of the amount of

the candidate molecular species if link *a* is correct and 0 if link *a* is incorrect, and S_p is a portion of the sum total activity in fraction *p* which is not accounted for by the activity of candidates assigned to that fraction.

16. A method according to any one of the preceding claims wherein the molecules are DNA.

17. A method according to claim 16 wherein information measured or determined for the molecules comprises partial sequence information.

18. A method according to claim 16 or claim 17 wherein information measured or determined for the molecules comprises fragment length following restriction enzyme digest.

19. A method according to any one of claims 16 to 18 wherein the activity measured is abundance.

20. A method according to claim 19 wherein abundance is measured as electrophoresis peak area for fragments of a given length.

Figure 1A

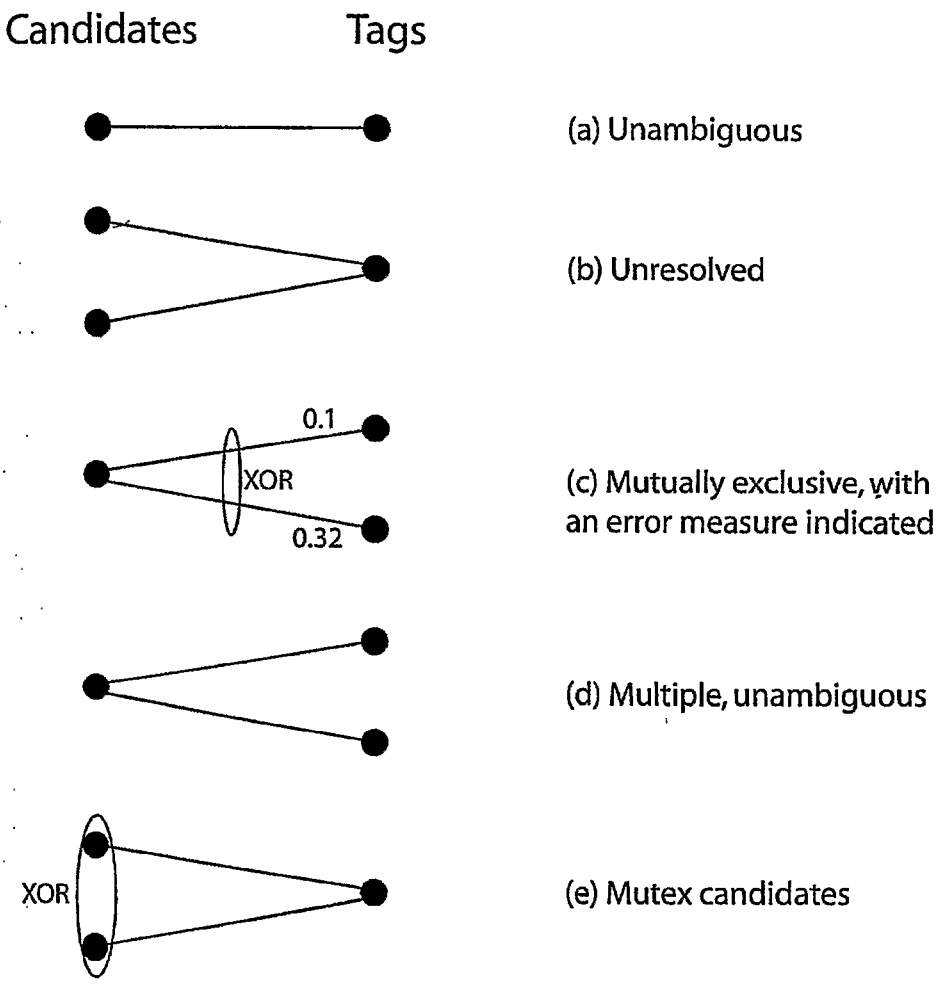
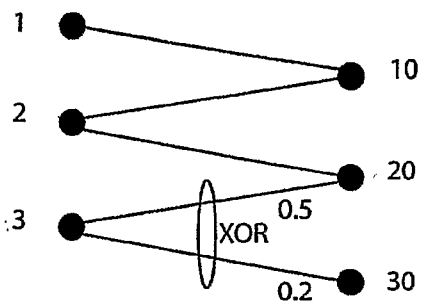
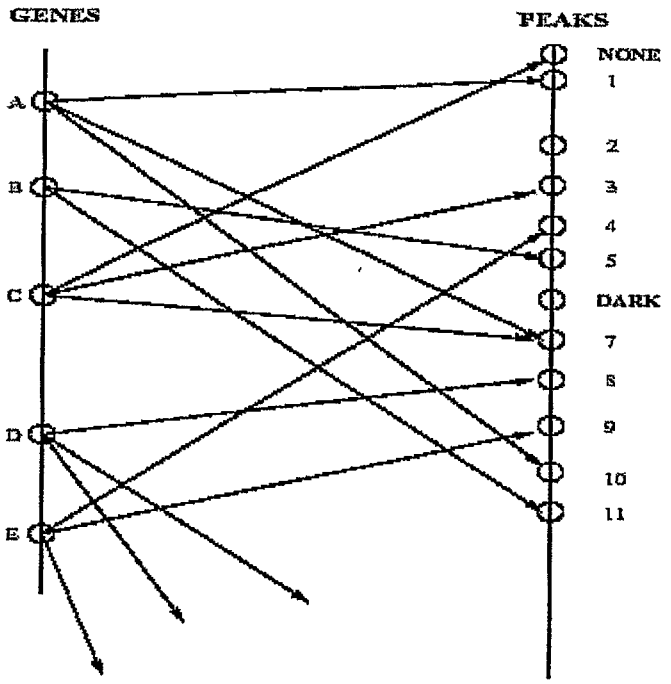


Figure 1B



A cluster of transcripts and tags interlinked in various possible ways. Tag-quantities are indicated next to tags, and transcripts are numbered. Error measures for mutex assignments are indicated on each link.

FIGURE 2



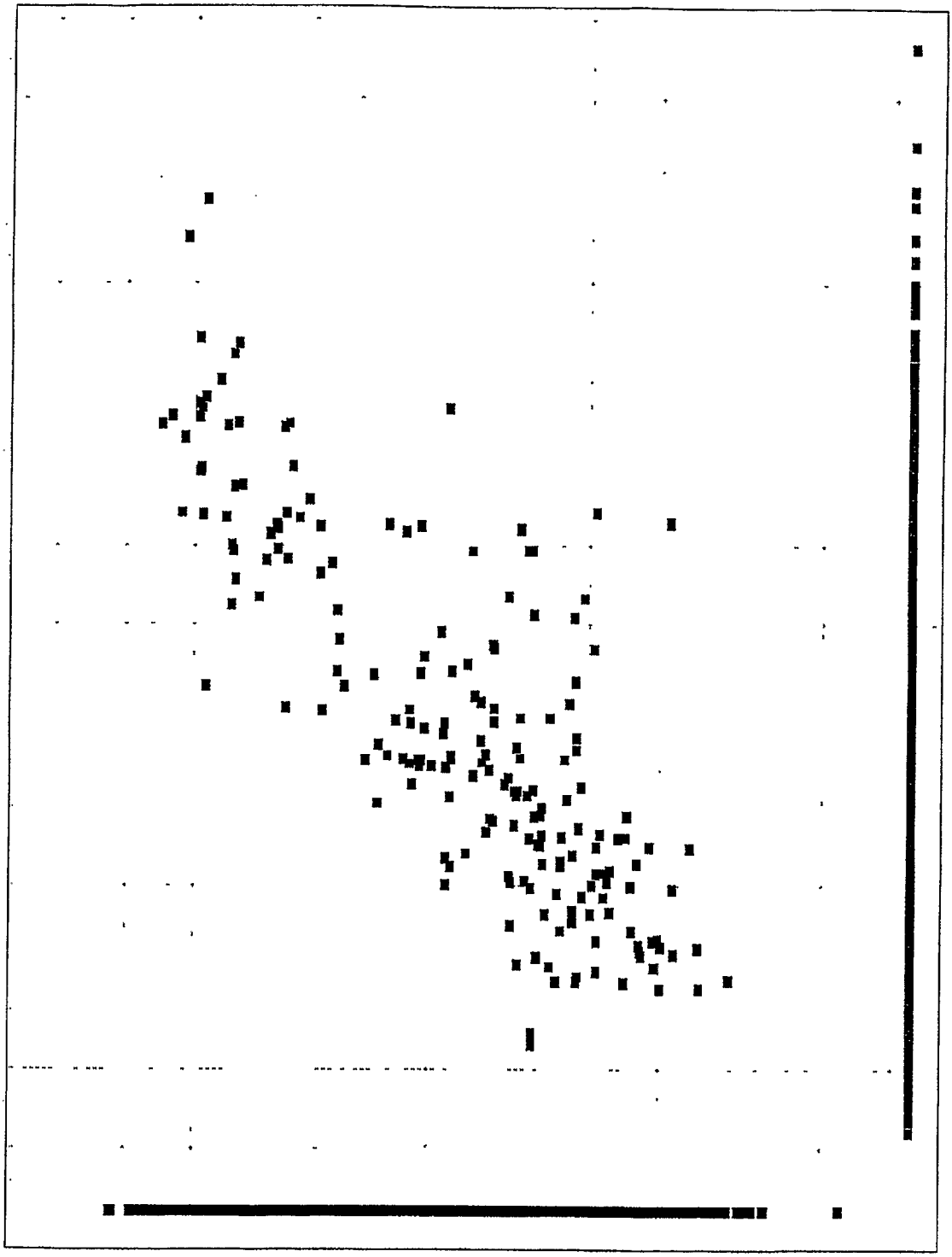


Figure 3