



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 148*

Software Tools for Design of Reagents for Multiplex Genetic Analyses

JOHAN STENBERG



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006

ISSN 1651-6206
ISBN 91-554-6551-X
urn:nbn:se:uu:diva-6832

Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Rudbecklaboratoriet, Uppsala, Tuesday, May 23, 2006 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Stenberg, J. 2006. Software Tools for Design of Reagents for Multiplex Genetic Analyses. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 148. 41 pp. Uppsala. ISBN 91-554-6551-X.

Methods using oligonucleotide probes are powerful tools for the analysis of nucleic acids. During recent years, many such methods have been developed that enable the simultaneous interrogation of multiple qualities of a sample. Many of these multiplexing techniques share common limitations. This thesis discusses new developments to overcome the problems of multiplex amplification of genomic sequences and design of sets of oligonucleotide probes for multiplex genetic analyses.

A novel molecular technique, termed the selector method, is described. This method allows circularization of an arbitrary selection of restriction fragments from genomic DNA, and the subsequent amplification of these circular products in parallel using common primers. The utility of the method is demonstrated by a 96-plex amplification experiment. Furthermore, the PieceMaker software for selection of restriction enzymes and restriction fragments is described. These two developments allow the selective amplification of subsets of genomes for further analyses.

A software tool for the design of sequence-tagged oligonucleotide probes is presented. The ProbeMaker software is a framework for design of sets of probes composed of separate functional elements, and uses an extension mechanism to incorporate support for new probe types as needed. An approach to a unified system for oligonucleotide design is also presented. This system will serve to decrease development times for new oligonucleotide design applications by allowing extensive code reuse.

Johan Stenberg, Department of Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, SE-75185 Uppsala, Sweden

© Johan Stenberg 2006

ISSN 1651-6206

ISBN 91-554-6551-X

urn:nbn:se:uu:diva-6832 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-6832>)

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

- I Fredrik Dahl, Mats Gullberg, **Johan Stenberg**, Ulf Landegren, Mats Nilsson (2005). Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.*, **33**, e71
- II **Johan Stenberg**, Fredrik Dahl, Ulf Landegren, Mats Nilsson (2005). PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res.*, **33**, e72
- III **Johan Stenberg**, Mats Nilsson, Ulf Landegren (2005). ProbeMaker: an extensible framework for design of sets of oligonucleotide probes. *BMC Bioinformatics*, **6**, 229
- IV **Johan Stenberg** (2006). Approaching a unified data processing model for oligonucleotide design. *Manuscript*

Contents

Introduction.....	9
Multiplex nucleic acid analyses	10
Aims of nucleic acid analyses	10
Genetic variation.....	10
Types of nucleic acid analyses	11
Mechanisms for probe-based nucleic acid analysis.....	12
Target recognition and sequence resolution	12
Amplification and detection	13
Coding and decoding of individual signals.....	15
Summary	16
Design of multiplex assays	17
Principles of multiplex oligonucleotide design	17
Different classes of design criteria	17
Design complexity	19
Sequence ranking and selection methods	21
Algorithms and software for oligonucleotide design	23
Melting point calculations	23
Existing software tools for oligonucleotide design.....	23
Present developments.....	26
Paper I. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments	26
Demonstration of the selector method.....	26
Comments	27
Recent developments	28
Paper II. PieceMaker: selection of DNA fragments for selector-guided multiplex amplification	28
The PieceMaker software	29
Results	29
Comments	29
Recent developments	30
Paper III. ProbeMaker: an extensible framework for design of sets of oligonucleotide probes	31
The ProbeMaker software.....	31
Comments	32

Paper IV. Approaching a unified data processing model for oligonucleotide design.....	32
The Comodo software system	33
Comments and perspectives	33
Concluding remarks	35
Acknowledgements.....	37
References.....	38

Abbreviations

cDNA	Complementary DNA
CNP	Copy number polymorphism
CNV	Copy number variation
C2CA	Circle to circle amplification
DNA	Deoxyribonucleic acid
MLPA	Multiplex ligation-dependent probe amplification
OLA	Oligonucleotide ligation assay
PCR	Polymerase chain reaction
RCA	Rolling circle amplification
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
T _m	Melting point
XML	Extensible markup language

Introduction

The amount of genomic information available today is enormous and growing quickly. More than 300 organisms have been completely sequenced, with draft sequences existing for many more¹. The human genome alone consists of more than 3,000,000,000 positions of information, many millions of which are known to commonly vary between individuals². Genomic data, in conjunction with tools for storing, integrating, searching, and analyzing the data, is an important resource for genetic analysis efforts with aims such as establishing the genetic background of disease, predicting how an individual will respond to drug treatment, or explaining the development of cancer. The availability of complete or partial sequences of many microorganisms is also an important resource for using genetic analysis techniques for purposes such as environmental monitoring, infection diagnostics, food safety, and for defense against bioterrorism and biowarfare.

To be able to utilize these resources, methods are needed to investigate the various types of genetic differences between individuals and between individual cells. Increasingly, methods are being developed that can be performed in multiplex, allowing many genetic qualities of a sample to be assessed simultaneously. Many of these methods use oligonucleotide probes to enable large-scale analyses of genetic or epigenetic variation or variation in gene expression levels. Although powerful, these methods share common limitations, including the requirement to carefully design the reagents. Some of these methods are also limited by problems associated with the amplification of many genomic sequences.

The work that is presented in this thesis aims to provide solutions to two of the problems associated with multiplex genetic analyses. The first problem to be attacked is that of simultaneously amplifying a large number of arbitrarily selected small sections of a genome. The other is that of selecting and designing the reagents required in present and emerging analysis methods. Before describing and discussing the present developments, current techniques for multiplex genetic analyses will be described, as will principles of computer-assisted design of oligonucleotide probes.

Multiplex nucleic acid analyses

The recent years have seen the development of numerous new methods for multiplex genetic analyses. These methods allow the simultaneous analysis of many nucleic acid targets in a sample in a single reaction. Multiplex assays have been applied for the analysis of several aspects and qualities of DNA or RNA target molecules.

The present work is concerned with the development of tools to support the establishment of new multiplex genetic assays. To give a background to this development, existing techniques for multiplex analyses will be briefly described in this section. This review is not intended to be complete, but rather to exemplify current methods. Examples have been selected that clearly describe the type of method or mechanism being discussed, but they are not necessarily the first instance of such methods or mechanisms. Some analytical techniques require the targets of the analysis to be amplified, and methods currently used to accomplish this will be described in some depth.

Aims of nucleic acid analyses

Investigations on nucleic acid molecules are carried out for a variety of reasons. Some investigations are explorative, such as those aimed at establishing the full genetic sequence of an organism. Others are of a more examining nature, including the diagnostics of known genetic variation and analyses to establish the type of microorganism present in a sample. Many methods have a comparative aim, such as finding genetic differences between individuals to establish the genetic cause of a particular phenotype, or to study evolution by the comparison of different species.

Genetic variation

The genomes of two human individuals are to a very large extent similar, but the existing genetic variation nevertheless has a large impact on the phenotype, including hereditary disease, response to viral or bacterial infection, and response to drug treatment. These variations can either be inherited, affecting all cells of an organism during its entire life span; or they can be acquired by the organism as an effect of DNA modification or damage by chemical agents, radiation or erroneous action of DNA

polymerases during DNA replication. Acquired mutations are present only in the affected cell and its descendants. They are thus generally not passed on to offspring, but may be involved in the transformation of normal cells to cancer cells.

Genetic variation comes in many different forms. The simplest form is the single nucleotide variation, involving the substitution of a single nucleotide for another. When the most common variant exists in less than 99 % of a population, this type of variation is referred to as a single nucleotide polymorphism, or SNP. Other variations on the sequence level include insertion or deletion of short sequences, different numbers of short repetitive sequence elements, and variation in the number of copies of longer sequences, including whole genes or parts of genes. The latter type of variation is often referred to as copy number variation (CNV), or copy number polymorphism (CNP) in the case of a variant present in more than 1 % of a population. Variations also exist that involve whole chromosomes or large parts of chromosomes, including the absence of chromosomes, the presence of extra copies of chromosomes, and the translocation of part of one chromosome onto another. Variation may exist also between different cells of the same individual. This also includes variation in modifications to the DNA sequence, such as the pattern of CpG methylation.

There are also many ways in which genetic variation can affect the phenotype. Variation may cause a gene product to lose its function or to gain another function, or it may cause the change of expression pattern of a gene so that it is expressed at other times, levels, or in other tissues than normal. A duplication or deletion event of a gene may also cause an abnormal amount of gene product to be produced.

Types of nucleic acid analyses

To examine the many types of genetic variation, there is a wide repertoire of tools available to the genetic analyst. These different methods generally adopt one of two strategies. One is concerned with establishing the presence, and sometimes also the amount, in absolute or relative terms, of a particular sequence of nucleotides in a sample. The other strategy has the objective of finding the sequence of nucleotides in the vicinity of another previously known, or assumed, sequence.

For a particular application, methods using either of the two categories may be applicable. For example, SNP genotyping is regularly performed using either of these types of methods³. For other purposes, a combination may be used. For example, the detection of a particular sequence in a sample may be sufficient to establish the presence of a family of microorganisms. To be able to discriminate closely related species, it may be necessary also to find the nature of adjacent sequence.

Methods employing both of the strategies commonly make use of oligonucleotide probes. The following section will describe reaction mechanisms commonly used for this and how they can be applied for different kinds of genetic analyses.

Mechanisms for probe-based nucleic acid analysis

Most techniques for nucleic acid analysis are based on the recognition of the target molecule by one or more oligonucleotide probes that base-pair to the target. Some methods rely solely on this hybridization, while many utilize one or more enzymes to process the hybridized probes, creating new molecular species in a target-dependent manner. In general, the products of these reactions contain some features that allow them to be amplified, identified, and/or visualized.

Probe-based analysis techniques generally consist of several reaction steps, such as amplification of the target sequence, hybridization of the oligonucleotide probes, enzymatic processing of probes, and decoding of individual signals. In some techniques, a set of steps are performed simultaneously and/or repeatedly.

Target recognition and sequence resolution

Target recognition is regularly based on the specificity of hybridization of oligonucleotide probes to their target nucleic acid molecules. A perfectly matched pair of sequences will hybridize more stably than will a pair with one or more mismatches. Thus, under suitable reaction conditions, the perfect match duplex will be more stable and therefore more abundant than the imperfect match. This property is utilized in allele-specific hybridization methods to detect the presence of particular sequences with a resolution of single base differences^{4,5}. The specificity of hybridization is generally not sufficient for this to perform well with complex samples such as human genomic DNA unless complexity reduction strategies are used^{6,7}, as will be described in the next section.

The ability of a ligase enzyme to discriminate single base differences can be utilized to achieve the single base resolution directly on genomic DNA. Methods using two recognition events coupled to a ligation step include the oligonucleotide ligation assay (OLA)⁸ and the related padlock probe assay⁹. In these methods, two probes hybridize adjacent to each other on the target and are joined by the ligation event. In the latter method, the probes are already connected at their distal ends, resulting in the formation of a circular molecule. By using two allele-specific probes that differ in a single position, the relative abundance of the two possible reaction products can be used to determine the genotype of the sample. Probes can also be hybridized a

distance apart on the genomic sequence. In this case, the resulting gap is first filled by the activity of a polymerase extending one of the probes until it can be ligated to the partner probe¹⁰.

Single-base resolution in sequence detection methods can also be provided by primer extension. To achieve this, allele-specific primers can be designed so that the 3' terminal nucleotide is matched to one or another variant of the sequence being analyzed. The formation of a primer extension product from a primer then indicates that the corresponding target variant was present¹¹. In a variant of this strategy, a blocked primer is used in conjunction with a proof-reading polymerase. This results in an extension product being formed only from mismatched primers¹². By using long probes immobilized on a solid phase in combination with stringent washes, even complex samples can be analyzed directly using single probes¹³.

Besides polymerization and ligation, so-called structure-specific endonucleolytic cleavage has been used for allele discrimination in SNP genotyping methods¹⁴. This utilizes the ability of certain enzymes to cleave a structure that is formed when two probes hybridize to a target strand in such a manner that the 3' end of one probe displaces part of the other probe.

Apart from its use for identifying sequence variants, primer extension is also the most commonly used reaction mechanism for determining the sequence adjacent to a known sequence. The Sanger sequencing method uses terminating nucleotides to halt the extension at different positions and resolve the sequence by gel electrophoresis¹⁵. The primer extension mechanism has also been put to use for SNP genotyping. The single-base extension, or minisequencing, technique uses the action of a polymerase to extend a primer by a single nucleotide and thereby determine the nucleotide present at the corresponding position in the sample¹⁶. The pyrosequencing method extends the primer one nucleotide at a time, registering the sequence of incorporated nucleotides, which can be used for short range sequencing as well as SNP genotyping¹⁷. These approaches require the target to be amplified, while the molecular inversion probe assay uses single-base primer extension in combination with ligation to achieve the specificity required for analysis directly on genomic DNA¹⁸. Again, by the use of long immobilized probes and stringent washes, the single-base extension method can be used directly on genomic DNA with a single probe¹⁹.

Amplification and detection

To report the results of a genetic analysis experiment, a detectable signal is required. This can be created by the attachment of detectable moieties such as fluorophores to the oligonucleotide probes or to the corresponding amplification products. Some methods use probes labeled directly with a detectable moiety^{4,20}, while for some methods that include a probe amplification step, the signal is attached to the amplification products based

on sequences included in the probes prior to the analysis^{18,21,22}. For other methods the signal is generated as part of the enzymatic probe processing step, as in Sanger sequencing¹⁵ and minisequencing¹⁶. Signals may also be created transiently during or after probe processing, *e.g.* by bioluminescence as in the pyrosequencing technique¹⁷. Several methods use a mass spectrometry read-out to separate different probes and reaction products, as reviewed by Tost and Gut²³.

Most genetic analysis techniques include an amplification step in order to obtain a level of signal that is sufficient for detection over any background. This may be performed by amplifying the target sequences or by amplifying the reacted probes. The first type of methods will increase the number of possible detection events, while the second type will generate a stronger signal from each detection event. Both methods involve the production of nucleic acid molecules. Signal amplification by other means may also be carried out at the time of detection.

Amplification of target molecules may also serve another purpose, that of reducing the complexity of the sample by directed amplification of the sequences that are targets of the analysis. Most methods that utilize a single recognition event on complex samples such as human genomic DNA require a complexity reduction step to achieve the desired specificity.

For random or undirected target amplification, various forms of whole genome amplification may be applied^{24,25}. Directed target amplification reactions are instead aimed at amplifying a subset of a genomic sample. Ideally, this subset should consist of all target sequences under investigation and no other sequences. Several methods have been proposed towards this end. The polymerase chain reaction (PCR)²⁶ has long been the method of choice for amplification of single sequences from genomic DNA, and one approach to amplify several fragments at once is to add multiple primer pairs to a reaction, possibly with 5' overhangs that incorporate common sequences into all amplification products for a subsequent round of amplification with a common primer pair. Although this strategy has been successful for low numbers of targets¹¹, the risk for amplification artifacts, such as truncated or false amplification products and primer dimers, increases with the number of targets approximately as $2n^2+n$ where n is the number of targets and thus primer pairs used²⁷.

Digestion of genomic samples with a restriction enzyme, followed by the ligation of common adaptor sequences to all ends, allows the amplification of many sequences at once using a common primer pair. By adjusting the PCR conditions it is possible to preferentially amplify fragments within a certain size range^{6,7}. This procedure will amplify a large number of fragments that are not of interest but it is useful as a complexity reduction method for some applications.

A hybrid between the abovementioned procedures is achieved by adaptor ligation and the use of one specific primer per target sequence and one

common primer. This requires a total of $n+1$ primers, substantially reducing the problems associated with ordinary multiplex PCR^{28,29}.

Callow *et al.*³⁰ present a method using two rounds of selective amplification with type II S restriction enzymes and specific adaptors, and demonstrate the method's ability to specifically amplify single fragments from genomic DNA. The method is multiplexable by the use of different adaptors, but it can be shown that it is not amenable to high levels of multiplexing because of limitations on which adaptors may be used together.

Methods that rely on probe amplification generally enabled this by the introduction of common sequence motifs into the probes. This motif is then used to perform a simultaneous amplification of all probes using PCR²², rolling-circle amplification (RCA)^{31,32} or circle-to-circle amplification (C2CA)³³ with common primers.

Coding and decoding of individual signals

When several targets are analyzed in parallel in a single probe-based reaction, it is usually necessary to encode into the probes what target sequence and/or sequence variant they are specific for, and then to decode this information after the detection has taken place, in order to separate signals originating from individual targets. Several methods exist for this coding and decoding that is the essence of multiplexing.

Coding is achieved by the introduction of a unique feature into every probe. This feature may be part of the nucleic acid, identifying the probe by its length^{20,21} or its sequence^{16,18,22,34}, but may also be some other feature, such as a fluorophore or a moiety of a specific mass. This coding feature may be introduced when manufacturing the probes or during the probe processing or probe amplification steps of the assay. To be able to encode more than one quality into a single probe, such as locus and variant for genotyping applications, more than one coding feature may be employed.

Decoding is performed based on the coding method used for the assay. With fluorescent coding, decoding is performed by fluorometric measurements, while mass spectrometry is used for mass-coded probes. Length-encoded probes are separated by electrophoresis. Probe encoded with specific sequences are generally sorted on an array of complementary oligonucleotide probes³⁵.

Not all methods for multiplex nucleic acid analyses encode target identity in probes in the sense described above. A complexity-reduced and labeled sample can be hybridized to a microarray with oligonucleotide probes complementary to all single nucleotide variations of a set of target sequences. The signal distribution over the array will reveal the relative number of each variant^{5-7,29}. A similar strategy can be employed for gene expression profiling³⁶. In these methods, the target identity is not encoded in the probes themselves, but in the location of the probe on the array.

Summary

Several techniques for multiplex genetic analyses have been developed using a number of reaction, detection, and identification mechanisms combined in various ways. Most mechanisms have found their use in many different methods. Furthermore, each of these methods has found several different applications.

As an example, a length-encoded version of the OLA, termed the multiplex ligation-dependent probe amplification (MLPA) technique, has been used for copy number analysis²¹ and CpG methylation analysis³⁷. A sequence-coded version in conjunction with microarray readout has also been used for profiling alternative splicing³⁴. The related padlock probe technique has also been used for SNP genotyping²², gene expression profiling³⁸, and the detection and identification of microorganisms³⁹.

As many of the increasing number of methods for multiplex nucleic acid analyses that are becoming available are variations on a number of themes, they share common properties and thus common limitations. In common for all probe-based methods is the need to design probes to be specific for their intended targets and to function well together if used in a multiplex assay format. This design is the topic of the next section.

Design of multiplex assays

This section will focus on the increasingly important problem of design of multiplex assays, including selection of target sequences and design of probes. Principles of probe design will be described and existing tools for computer-assisted oligonucleotide design for multiplex assays will be briefly reviewed.

In the following discussion the term *genomic sequence* will be used for sequences that are of genomic origin, including RNA transcripts and cDNA as well as amplification products. The term *target sequence* is used for a sequence that is a subsequence of some genomic sequence of interest, and to which a probe will be designed to base-pair. A *probe sequence* is a sequence that base-pairs to a target sequence, and may include other sequences to be used for probe processing or identification, as described in the previous section.

Principles of multiplex oligonucleotide design

Once the aim of a particular investigation has been established, in the sense that *e.g.* a number of SNPs for a genotyping project or a number of genomic regions for a resequencing project have been selected, the task is to find the actual set of target sequences for which to design the set of probes. For a single probe, the problem of target selection involves the identification of a subsequence of a given genomic sequence to which the probe should base-pair, subject to given selection criteria. The nature and complexity of the target sequence selection problem depend on the application at hand as described in the following examples.

Different classes of design criteria

Let us first consider the selection of a single probe for hybridization to a specific nucleic acid, such as a particular gene or part of a gene. The target sequence selection problem is here to select a subsequence of this longer genomic sequence, and the number of possible choices is determined by the length of the sequence in conjunction with design criteria about the desired length or melting point, T_m , of the probe sequence. With these considerations in mind, every possibility can be evaluated and a target sequence selected.

Two classes of such design criteria may be used in this selection. Class I criteria apply to the properties of the target sequence itself, such as length, hybridization stability, and potential for homodimer formation, while class II criteria concern the target sequence in relation to the genomic sequence context, mainly the ability of the resulting probe to uniquely identify the target sequence among all other genomic sequences present in the sample.

As another example, consider selecting a pair of PCR primers to amplify a certain genomic sequence. The problem posed here is similar, but requires two sequences to be selected, thus introducing a second dimension to the problem. The number of possible choices now depends on the number of possible choices for each primer and limitations on the distance between the two primer sequences. A new class of criteria (class III) is necessary to avoid selecting primer sequences that work poorly together, e.g. by forming primer dimers.

For SNP genotyping assays such as minisequencing, the target selection problem is simpler and consists of choosing one or the other strand to use as target sequence. If both strands are to be analyzed, the solution to the target selection problem is reduced to choosing the length of each target sequence to find a suitable T_m .

When selecting targets for multiplex assays, a further class of criteria (class IV) applies in order to avoid using incompatible target sequences in the same assay, such as pairs of sequences that have a large degree of complementarity that could inhibit the probing reaction or produce false reaction products.

To summarize, four classes of design criteria can be identified for target sequences.

- I Criteria for the target sequence itself
- II Criteria for the target sequence in relation to the genomic sequence context
- III Criteria for the target sequence within the set of target sequences for the same genomic sequence
- IV Criteria for the target sequence in relation to target sequences for other genomic sequences

When a set of target sequences has been selected, it is time to design a set of probes for these sequences. The probe sequence design problem is to find one or more probe sequences for each target, such that the probes satisfy probe design criteria. In addition to designing probes for all targets, designing probe sets for multiplex assays also includes trying to predict how probes will work together and selecting probes that minimize problems due to unintended interactions between different probes or between probes and targets. The same four classes of constraints apply also for probe design.

- I Criteria for the probe sequence itself
- II Criteria for the probe sequence in relation to the genomic sequence context
- III Criteria for the probe sequence in relation to other probe sequences for the same target
- IV Criteria for the probe sequence in relation to probe sequences for other targets

In some cases, probe design is a trivial task, completely determined by the target selection. This is the case for regular PCR, where the probes only contain sequences that are complementary to the target sequence. For more complex probes such as tag-coupled minisequencing primers, OLA probe pairs or padlock probes that contain primer and tag sequences, the probe design problem is a complex problem, including the evaluation of many candidate probe sequences, in a manner similar to that described for target sequences.

Design complexity

With the descriptions of the target sequence selection and probe sequence design problems presented above, it is clear that the requirements for design algorithms differ widely between applications. For methods such as regular PCR, only target selection needs to be performed, while genotyping methods that are performed on PCR-amplified material, such as tagged minisequencing, may require two rounds of target selection (selection of PCR primers to amplify sequences containing the SNPs, and selection of strand to genotype) and one round of probe design (selection of tags for each minisequencing primer). The complexity of these problems for a given application is determined in part by the number of possible choices of target sequences or probe sequences. Again, some examples will be used to illustrate this.

Consider as example A the design of a single PCR primer pair to amplify a genomic sequence. Let the requirements on the length of the amplified sequence be such that the number of possible positions of one primer is 100, and that for each choice of this primer, there are 100 possible positions for the second primer. For simplicity, consider there to be a single valid choice of primer length for each primer position. There are then 10,000 possible choices for this primer pair.

Consider as example B the design of 10 such primer pairs for use in the same reaction. There are still 10,000 possible primer choices per genomic sequence to be amplified, and if the different selections were independent problems, there would be a maximum of 100,000 primer pair choices to evaluate. However, since the primer pairs are to be used together, these

problems are not independent and the number of possible combinations is $10,000^{10}$, or 10^{40} .

Consider as example C the design of a padlock probe for a specific genomic sequence. Let the target selection be limited to a single possible target sequence. The probe should include a tag sequence for array hybridization, selected from a library of 10 such sequences. There are thus 10 possible probe candidates.

Consider as example D the same design as in example C, but with the possibility to choose either strand to hybridize the probe to. There are then 2 possible combinations of target sequences. For each target sequence there are still 10 possible probes, making the total number of possibilities 20.

Consider as example E the design of 10 such probes for multiplex use. Each probe should be equipped with a unique tag. Even if class III and IV probe design criteria are disregarded, design of each probe cannot be considered to be an independent problem since the use of a tag in one probe prohibits its use in other probes. Thus, the number of possible combinations of probes and tags is $10!$, or approximately $3.6 \cdot 10^6$, if only combinations where all probes are successfully designed are considered. Considering that probes may fail if no tag selection satisfies the probe design criteria, there are 10^{10} possible combinations. This is true for every possible combination of target sequences, giving a total number of $\sim 10^{13}$ possible designs.

Let n be the number of probes to be designed, t the number of possible selections of target sequence for each probe, and p the number of alternative probes possible for each probe. The total number of combinations for the above examples can then be expressed as $t^n \cdot p^n$, or $(t \cdot p)^n$, as summarized in table 1.

<i>Example</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>t</i>	10,000	10,000	1	2	2
<i>p</i>	1	1	10	10	10
<i>n</i>	1	10	1	1	10
$(t \cdot p)^n$	10,000	10^{40}	10	20	$\sim 10^{13}$

Table 1. Number of possible combinations for different design examples

From these examples, two things are evident. Firstly, the complexity of the multiplex oligonucleotide design task grows quickly with increasing numbers of possible target and probe sequences, and requires computer software to be approachable. Secondly, even with the use of software for computer-assisted design, the large number of possibilities to evaluate prohibits exhaustive selection methods for most design problems.

Sequence ranking and selection methods

To be able to select among the different target and probe sequences possible for a given application, a ranking system is required. This way, each combination can be evaluated and assigned a rank, which can be used to select among combinations. This ranking system must have at least two levels, allowing the target or probe sequence to be either accepted or rejected. A finer ranking system allows the identification of target or probe candidates that are sub-optimal, but considered to be acceptable in spite of some problems.

An ideal selection of target and probe sequences is one where all targets and probes satisfy all the design criteria in use, such that all selected targets and probes are ranked at the top level. If such an ideal combination exists, the selection procedure can safely be stopped once that combination has been found. Often, however, there is no combination of targets and probes that satisfies all design criteria. In such cases, all possible combinations have to be evaluated to ensure that an optimal one is found. A definition of an optimal combination can be a combination that has the highest number of probes/targets of the highest rank of all possible combinations, or one which maximizes the sum of ranks for all targets/probes. For a simple two-level ranking system these two definitions of optimality become equal.

An exhaustive method allows and requires the evaluation of all possible combinations, in the worst case. As discussed in the previous section, exhaustive methods become complex when the number of probes increases. Other methods are thus required to find good target and probe combinations. By evaluating all components of a target or probe selection separately before combining them, the complexity can be reduced.

Consider design example A (design of a single PCR primer pair) from the previous section, using a two-level ranking system. Criteria of classes I, II, and III apply. By evaluating all possible selections of primer 1 and primer 2 separately, using class I and II criteria, the number of acceptable primer combinations is reduced, as only acceptable selections of each primer need to be tested together. In fact, the first pair of two acceptable primers that also satisfies the class III criteria is an optimal primer pair, assuming that primers of the lower rank are discarded. If a multi-level ranking system is used, the problem becomes more complex again, and the number of combinations that have to be evaluated depends on what emphasis the ranking system places on the different classes of criteria, and what optimality definition is used. In particular, an optimal primer pair does not necessarily consist of primers of the highest rank. Similarly, an optimal selection of ten primer pairs for example B does not necessarily consist of ten primer pairs that are optimal for their individual selection problems.

An approach that significantly reduces the amount of evaluations performed is the greedy selection method. A greedy algorithm can be

described as one that tries to find a solution to a problem by finding an optimal solution to each of a series of sub-problems. This is applicable to the problems of target and probe selection, at the cost of not always finding an optimal solution. For example A, a greedy approach would be to first find an optimal selection of one primer, and with this done, find the best selection of the second primer, using all classes of criteria. The maximum number of combinations to evaluate is reduced to 200. For example B, a greedy approach would be to select each of the 10 primer pairs in sequence. Each primer pair could be selected either exhaustively or greedily, yielding 100,000 or 2,000 possible combinations to evaluate, respectively, in the worst case. Table 2 summarizes the effects on the design examples of using greedy methods for selection of target sequences and probe sequences, for target and probe sequence combinations, and for both.

<i>Individual</i>	<i>Combination</i>	<i>A</i> [†]	<i>B</i>	<i>C</i> [†]	<i>D</i> [†]	<i>E</i>
Exhaustive	Exhaustive	10,000	10 ⁴⁰	10	20	10 ¹³
Exhaustive	Greedy	10,000	100,000	10	20	200
Greedy	Greedy	200	2,000	10	12 [‡]	120 [‡]

[†] These examples do not include combination selection.

[‡] In this case, the greedy individual selection approach is to first select target strand (2 possibilities), then select tag (10 possibilities).

Table 2. Number of possible combinations for different design examples using greedy and exhaustive methods for selection of individual target and probe sequences (or pairs of sequences) and of combinations of these sequences.

Clearly, greedy strategies significantly reduce the complexity of the design task. This is particularly so when evaluating target criteria of class IV, which require comparing target or probe sequences to all other sequences in the set being designed. These steps are generally the most time-consuming, and reducing the number of such comparisons greatly reduces the time required for a selection task.

The choice of selection method for a particular application is governed by the requirements on finding an optimal solution, and the time available for the design. For some applications, it may be imperative that the design succeeds for all targets, while other application may allow some designs to fail, either by using a redundant set of targets or by being able to use substitute targets. The latter may be performed by the use of an iterative design scheme, where targets for which the design fails are removed from the set and substitute targets added and the design reiterated with the already selected targets and/or probes in mind. An iterative design may also be used to redesign probes for failed targets using less stringent design criteria. Another approach to rescuing failed targets is to divide the set of probes into subsets, where each subset contains probes that work well together.

Algorithms and software for oligonucleotide design

Numerous software tools exist to aid in the design of oligonucleotides. The most common tasks performed by these tools are the selection of extension primers for various applications such as PCR, genotyping, or sequencing⁴⁰⁻⁴⁴, and the selection of oligonucleotide probes for genomic DNA or cDNA microarray hybridization experiments⁴⁵⁻⁴⁹. These available tools define a number of selection criteria for oligonucleotides and various algorithms to evaluate probes or primers based on these criteria. This section will describe some of the criteria that are commonly used.

Melting point calculations

Oligonucleotide probes recognize targets based on their sequence by the formation of a probe-target hybrid duplex. The stability of a duplex is often described by its melting point (T_m), which is defined as the temperature at which half of the potential duplexes are formed. Some different models exist to calculate the T_m for a duplex of known sequence, and will be briefly reviewed below.

Factors influencing the stability of duplexes include the number of hydrogen bonds as determined by the number of AT and GC base-pairs in the duplex, base stacking between adjacent base pairs, the presence of cations such as sodium and magnesium ions, and the presence of denaturing agents such as formamide and dimethyl sulfoxide. The simplest models address only the contributions from hydrogen-bonding, while more complex models address several of these factors. Wetmur⁵⁰ describes many aspects of probe hybridization, while von Ahsen *et al.*⁵¹ provide a brief compilation of different approximate models. The most accurate predictor of T_m is the nearest-neighbor model as described by Breslauer *et al.*⁵². See also Owczarzy *et al.*⁵³ and SantaLucia and Hicks⁵⁴ for a detailed description and review of the nearest-neighbor model. The nearest-neighbor model may also be applied to duplexes that are not perfectly matched. Adjustments to the T_m for some reaction components often present during enzymatic processing of probes are described by von Ahsen *et al.*⁵¹

Existing software tools for oligonucleotide design

There are two dominating ranking and selection strategies among existing software tools. The first strategy is that of successive filters^{43,47,48}. In this strategy, each target or probe sequence candidate is subjected to a series of consecutive tests, each with two possible outcomes, accepted or rejected. Only candidates that are accepted by one filter are passed to the next. Any candidates that pass all filters are considered acceptable for use. With the other strategy^{42,44}, a set of tests are carried out on a candidate, and a rank is

calculated based on the results of all tests, weighting the results of the different tests in some fashion. This rank can then be used to filter, sort or select candidates.

By ordering the tests according to the required execution time, the former strategy of successive filters allows time consuming tests to be avoided for candidates that fail in earlier tests. The second strategy allows the selection of sub-optimal probes or combination of probes, something which is often necessary for complex design problems where no optimal solutions exist. Another advantage of the second strategy is that it is generally easier for the user to find out what are the main reasons for design failure. Some programs use a combination of the two strategies^{46,49}, applying the former strategy for some classes of design criteria, and the latter strategy for some.

The most common criteria used to evaluate probes for hybridization-based assays is the presence of certain undesirable sequences, the risk for secondary structure formation or homo- and heterodimer formation between probes, and the risk for cross-hybridization of probes to other targets. For polymerase-based assays also false priming and primer-dimer formation is checked for. Commonly used design criteria are summarized in table 3.

Criterion class	Examples
Class I	Probe-target T_m
	Target sequence length
	Probe secondary structure (hairpin formation)
	Probe homodimer formation
	Probe self-priming (for polymerase-based applications)
Class II	Specificity (hybridization to correct target only)
	False priming on genomic sequence
Class III	Probe heterodimer formation
	False priming on other probes
Class IV	Probe heterodimer formation
	False priming on other probes
	Probe-tag cross-hybridization (for tag-based readout)

Table 3. Criteria commonly used in existing oligonucleotide design software

Most of the algorithms used to implement the evaluation of these criteria are concerned with the comparison of sequences to find regions of similarity or complementarity. These algorithms reach from simple string comparisons to dynamic programming algorithms for constructing alignments based on nearest-neighbor thermodynamic parameters⁴². The more complex algorithms yield better predictions to the cost of being more computationally expensive.

Many of the available primer design tools do not support the design of multiple primer pairs at once, and those that do generally do not use target criteria of class IV, effectively making the design problems independent. Furthermore, primer and microarray probe design software are generally concerned only with target sequence criteria. One exception to this is the SBEprimer software⁴². This program allows the design of sets of sequence-tagged primers for multiplex minisequencing, and evaluates the different classes of probe sequence criteria.

Present developments

The papers that constitute this thesis describe two distinct developments. Papers I and II demonstrate the so-called selector technique, a novel method for selective amplification of genomic sequences, and a software tool that solves the task of determining the reagents required for applications of this new method. Paper III presents an extensible software framework for design of sets of oligonucleotide probes consisting of multiple sequence elements. Paper IV describes the continuation of this work towards the development of a unified system for oligonucleotide design.

Paper I. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments

As discussed elsewhere in this thesis, current strategies for multiplex amplification of target sequences are limited in the level of multiplexing that can be achieved, or generate undesired amplification products. In this paper we describe and demonstrate a novel method for selective amplification of many genomic DNA sequences in parallel.

A genomic DNA sample is digested with one or more restriction enzymes, forming a mixture of different fragments. After denaturing the digestion products, the single-stranded fragments that contain sequences of interest are circularized by hybridization and ligation guided by partially double-stranded oligonucleotide constructs called selectors. The selectors consist of a common central part that is double-stranded and 10-25 nucleotides long sequences at the 5' and 3' ends that are single-stranded and that are complementary to the ends of the targeted restriction fragments. By utilizing an enzyme capable of structure-specific endonucleolytic cleavage^{55,56} it is possible to cleave off a 5' part of the fragment prior to the ligation, thus circularizing and thereby selecting a sub-fragment.

Demonstration of the selector method

To demonstrate the multiplexing ability of the selector method, we selected a set of 96 clones from a 7500 clone cDNA microarray intended for gene expression studies. The genomic sequence corresponding to each clone was

found using Blast⁵⁷. Selectors were then designed to amplify genomic sequences that would hybridize to the selected clones on the array. In this design we required fragments to be between 140 and 160 nucleotides in length, and to contain at least 70 bases of sequence complementary to the cDNA on the array. The selector design was carried out using the PieceMaker software, described in paper II. Two parallel digestion reactions, each with two restriction enzymes, were required to obtain suitable fragments for all targets.

DNA samples were restriction digested in two parallel reactions, circularized and then combined and amplified in multiplex by PCR. The fluorescently labeled amplification products were then hybridized to the cDNA array. Seven of the 96 cDNA clones turned out to be absent from the microarrays. Of the 89 fragments specific for the remaining clones, we successfully circularized, amplified, and detected 79, or 89 %, in at least three out of five replicate experiments, while 71 (80 %) were detected in all experiments. Experiments using genomic DNA from different individuals showed that the inter-individual variation was similar to the intra-individual one.

To investigate if further multiplexing of the selector method will be limited by the ability to select combinations of restriction reactions, we selected five sets of arbitrary genomic sequences and performed *in silico* digestion, fragment selection and selection of reaction combinations using the PieceMaker software. The results indicate that the design success rate will not decrease with an increasing number of target sequences.

Comments

The selector method compares favorably with other methods for directed target amplification. The specific selection of fragments is made possible by the requirement for two hybridization events, and the circularization of selected fragments enables the distinction of correct products from any erroneous reaction products, such as fragments that have had different selectors attached at each end. The fact that selectors are used at concentrations much lower than *e.g.* PCR primers should also substantially reduce the risk for unintended amplification products.

The selector method has the potential of being used as a preparation step for assays where the specific amplification of an arbitrary selection of sequences is required. One such application is the parallel resequencing of large numbers of DNA molecules, an area where considerable effort is being spent^{58,59}. This would take advantage of the full potential of the selector method by using the full information content of the amplification products.

From another perspective, the selector method is a multiplex probing technique that extracts parts of the genome into the processed probes. Several probe-based methods use a single primer pair to amplify all reacted

probes. For these methods, products from the amplification of unreacted probes cannot always be discriminated from desired reaction products, resulting in a high level of background signal. The selector method is expected to have fewer problems with background, since an amplified unreacted probe would not contain the genomic sequence.

Recent developments

Since the publication of paper I, we have begun applying the selector method in a number of applications, including for parallel resequencing, sequence copy number measurements, and DNA methylation analysis.

The resequencing approach involves using selectors to amplify a set of exon sequences and to further process the amplification product in a micro-emulsion PCR followed by sequencing using a high-throughput solid phase pyrosequencing platform as previously described⁵⁹. The initial results are promising, but one remaining issue is to achieve an even level of amplification for all targets (Dahl, *et al.* work in progress).

By selecting fragments of different lengths for each target, the corresponding amplification products will likewise be of different lengths. Using fluorescently labeled primers in the amplification step allows the separation and detection of products from each target by electrophoresis. The signal intensity for each probe will reflect the number of targets that were present in the sample. By comparing a sample to a reference it is possible to establish the relative copy numbers of the target sequences. This is similar to the previously described MLPA technique, which has been applied for measurements of aneuploidy and gene dosage, as well as for CpG methylation analysis. An advantage of the selector technique is that the probes are easier to produce than MLPA probes that have lengths of up to a few hundred base pairs. Our initial results show that the selector method can be used for high-precision gene copy number measurements of tens of targets in multiplex (Isaksson *et al.*, in preparation). By using methylation-sensitive restriction enzymes, this technique can also be extended to epigenetic studies (Isaksson *et al.* work in progress).

Paper II. PieceMaker: selection of DNA fragments for selector-guided multiplex amplification

During the planning of the experiments to demonstrate the multiplexing capability of the selector method it became evident that the assistance of a computer program was required to perform the reagent design. The PieceMaker software was developed to support the selection of restriction enzymes to generate restriction fragments for selector-guided amplification.

The PieceMaker software

PieceMaker is an interactive, graphical program that allows the user to perform *in silico* digestion of DNA sequences, evaluate the resulting fragments and select an optimal combination of restriction enzymes. An optimal combination is one that maximizes the number of targets for which suitable fragments are created in a given number of parallel digestion reactions.

A set of target sequences is digested *in silico* with a number of restriction reactions, each using one or more restriction enzymes, and taking into account uncertain presence of cleavage sites due to known variation in the input sequences. Restriction fragments are evaluated by filtering based on fragment length, content of target sequence, and optionally the existence of a suitable site for structure-specific cleavage.

An exhaustive approach is used to find an optimal solution to a given selector experiment design problem. To support the selection of restriction reactions for different types of applications, two selection schemes are available. One scheme requires at least one suitable fragment per target, while the other scheme requires that the whole target region is included in one or more selectable fragments. The former scheme is useful for finding fragments for copy-number measurements, where one marker per target is sufficient, while the other scheme is intended to be used for resequencing applications, where the objective is to sequence the complete target.

Results

To show the utility of the PieceMaker software and to investigate the ability to design selectors for actual applications, we performed two designs, one aimed at amplifying target sequences for genotyping of 100 randomly chosen SNPs, the other aimed at amplifying the complete sequences of 101 exons from three different genes. The results indicate that the applicability of the selector method to such ends will not be limited by design. The choice of fragment evaluation parameters, such as fragment length range and the maximum length of the flap (the sub-fragment cleaved off by the structure-specific cleavage), has a large impact on the design success rate. To begin to establish some design rules, we tried a few different designs experimentally. We have thus circularized fragments from 100 to 1000 nucleotides in length and used flap lengths of up to 1000 nucleotides.

Comments

Three types of considerations have to be made when designing selectors for any particular application. The reaction mechanisms involved in the selection reaction impose limits on fragment lengths. Very short fragments

will not be circularizable due to the rigidity of the DNA molecule. Very long fragments will form circles less efficiently due to the long distance between the two ends. Also, long fragments may be more difficult to denature which will adversely affect the circularization. Furthermore, the properties of the amplification method (PCR) may impose further limits such as the GC content of the selected fragment. There are also some limitations imposed by the intended application. For a resequencing application, the fragments should not be longer than what is possible to sequence, while for length-encoded multiplexing they should be within a size range suitable for the instrument on which they will be separated. Finally, the nature of the target sequences themselves will be limiting. For example, the presence of repeated sequence elements caused the design of one of the example targets to fail regardless of the design criteria used.

The PieceMaker program has served well in the design of several selector experiments, including the applications described under paper I, above. As more sets of selectors are designed, we will learn more about how to design probes that will function well. This knowledge can then be incorporated into the design software. There is also potential for other improvements to the usability and performance of the software. It would be beneficial if the software could import restriction enzyme data directly from sources such as REBASE⁶⁰, and also to import restriction maps that are generated by other tools. Also, as the number of designs grow, it could be convenient to store restriction mapping results in databases to reduce the time required to perform new designs on the same target sequences.

Recent developments

Since the publication of paper II, the PieceMaker software has been developed continually. It is now possible to specify the minimum distance from the position of the structure-specific cleavage at which known variations are allowed to occur. This should decrease the risk of losing individual amplicons due to poor efficiency of hybridization or ligation caused by sequence variability. Also, the fragment evaluation and reaction combination selection tasks have been delegated to plug-in software components. This has simplified the addition of new design criteria. Furthermore, performance improvements have been made that significantly reduce the time required for *in silico* digestion.

Paper III. ProbeMaker: an extensible framework for design of sets of oligonucleotide probes

As discussed earlier, the design of oligonucleotide probes for multiplex genetic analyses requires the aid of computer software to be tractable. For probes consisting of multiple sequence elements, such as padlock probes²² and the selector probes described in paper I, such software has not been publicly available. This prompted us to develop the ProbeMaker software. As there are many types of applications that utilize this type of probes, our objective has been to design a flexible and extensible software framework for multiple types of designs.

The ProbeMaker software

The design task performed by ProbeMaker is defined by a set of target sequences, sets of tag sequences and corresponding tag selection modes, design parameters, a set of analyses, and an acceptance and selection scheme.

The target sequences define templates for which up to two complementary sequences will be constructed. For each target, one probe is designed that consists of sequence elements complementary to the templates defined by the target and one sequence from each of the sets of tag sequences. The lengths of the two target-complementary sequences are determined based on design parameters for length and T_m , as calculated by a nearest-neighbor model.

The so-called 'tag' sequences, which may have different purposes, such as to allow amplification, restriction digestion or hybridization, are allocated to the probe based on a selection mode specified for each set of tags. This makes it possible to decide whether tags should be present in all individual probes, unique for each target-specific probe, or *e.g.* distinct for the allele-specific probes at any locus. Depending on the tag sets and selection modes, many possible combinations of tags may be possible for a given probe and are selected among by generating, testing and evaluating probe candidates.

Candidate testing is performed by the specified analysis modules, which define various criteria. The test results determine if candidates are acceptable and which candidate should be used, as defined by the acceptance and selection scheme in use.

ProbeMaker is equipped with an extension mechanism that facilitates the addition of new functionality, such as new target types, probe analyses, and acceptance and selection schemes.

To cope with the large numbers of probe candidates and combinations of probes for large number of targets, the tag selection is performed using a greedy method with an exhaustive strategy available for the sub-problem of

selecting pairs of allele-specific probes. Thus, ProbeMaker will not always find an optimal solution to the design problem.

By default, the program uses a three-level ranking system, but this can be changed by providing new acceptance and selection schemes. ProbeMaker is capable of selecting probes based on the four classes of probe sequence criteria, and it can also evaluate target sequence based on all four classes of target sequence criteria. However, although rudimentary selection for small numbers of possible targets (such as choosing one or the other strand of a sequence) is possible, ProbeMaker is not suited for target sequence selection. This may require other software, such as PieceMaker, presented in paper II.

Comments

For historical reasons, we initially aimed at creating a tool for the computer-assisted design of padlock probes. The many possible applications and thus different designs of padlock probes, in conjunction with the fact that optimal parameter values for probe design are not yet known, prompted us to construct a piece of software that was flexible in allowing different types of target sequences, different compositions of the functional elements of probes, and different probe evaluation criteria. In spite of this, the early versions of the program were still limited to padlock probe design. As the design of selectors was later required, we redesigned ProbeMaker to solve also this task, and in the process of doing this we added the extension mechanism described above to support present and future types of probes.

The software has enabled us to swiftly design probe sets for detection of bacteria (Jarvius *et al.*, submitted) and viruses (Banér *et al.*, submitted). With the capabilities provided by the extension mechanism, we have also used the software to design sets of selectors, including those used in paper I, and to design probes for proximity-ligation⁶¹ based characterization of DNA-protein interaction (Gustafsdottir *et al.*, in preparation).

Paper IV. Approaching a unified data processing model for oligonucleotide design

The large number of oligonucleotide design programs already available illustrates the increasing usefulness of such tools, and it also demonstrates that much effort has been put into software development in this area. Although many functions are similar or equivalent in these different programs, this area of software development has a low level of code reuse. In paper IV, I describe a prototype for a unified oligonucleotide probe design software system. The aim of this software project is to provide users with the

possibility to create design applications that are tailored to their specific requirements by combining existing components and services.

Current programs define a number of probe evaluation criteria and algorithms with which to evaluate candidate sequences. The software project presented in paper IV intends to provide a framework in which to combine these to build new applications that support the development of new analysis methods.

The Comodo software system

The software system, named 'Common models for design of oligo-nucleotides' or Comodo for short, consists of several parts. A data model describes the types of entities involved in probe design, and has been constructed to allow the representations of many types of probes through a generic set of data structures. Also, the model may be extended with new entities if need be. Individual data objects representing targets and probes are generated and processed as dictated by a design scheme.

A design scheme consists of interconnected software components that fill different roles of the design process, such as candidate generation, analysis, and filtering. The details of the actions taken by each component are determined by the services used by that component. A service is an independent software unit that performs an atomic task, such as calculating the T_m of a probe-target hybrid, or sorting probe candidates based on a rank value.

Design schemes are created programmatically or by manually editing a design scheme file. A design job is performed by submitting a design scheme and a set of input data to the enactor, or design scheme execution engine. The enactor will process the input data according to the design scheme specification, and return the resulting output data to a user or to a client program.

In order to demonstrate the system two design schemes were created, one for designing microarray probes for expression profiling, and one for designing padlock probes for bacterial detection. The schemes were executed with results comparable to those of other software.

Comments and perspectives

For the Comodo system to become really useful, it must fulfill a number of criteria. It must be possible for users to create and add new functions to the system, and to adapt it to their specific needs. Therefore, the source code will be distributed for free under a free software license. This, along with the choice of Java as the implementation language, will make the potential user base large, as the system will run on most available computer systems and there will be no license cost attached with either use or further development.

Secondly, to attract a large user base, the system needs to be applicable for different categories of users. This ranges from basic users who use the system for designs on a routine basis, to advanced users who contribute to the development of the system. Accordingly, the design scheme model can be utilized in several ways. A software application can create a design scheme programmatically in such a way that it is hidden from the user. This utilizes only a small part of the system's potential, but still makes it easier for the application developer to make changes to the system, and to reuse previously developed components. By defining the design scheme as an XML file that is read into the program on every run, users with understanding about the scheme have the possibility to change the order and nature of the processing steps, while still essentially hiding the complexity from those that so desire. More advanced or experimentally inclined users may wish to have a graphical editor with which to compose entirely new design schemes.

The system should also support the integration of design processes with other upstream and downstream systems. This is facilitated by allowing the use of an existing data model and XML-based file format for transmitting input and output data. A design process can easily be integrated with work flows built in the Taverna workbench⁶², and could also be included as a component of such a workflow.

Comodo can be run either as a standalone program on the local machine, or as a service located on a remote machine. The performance of the software could be increased by adapting it for parallel or concurrent processing and running it on a multiprocessor system.

Concluding remarks

Multiplexed nucleic acid analysis techniques are and will increasingly be important tools for gaining knowledge about genetic disease mechanisms. Many oligonucleotide probing methods have been developed for the purpose of performing SNP genotyping. As other types of variation now come into focus, these methods will be adapted to new applications, including the analysis of copy number variations and methylation. Methods will also be put to use for the detection and identification of microorganisms for purposes such as food and water safety, environmental control, and defense against biowarfare or bioterrorism. In these settings there will be room for methods able to perform at moderate as well as very high levels of multiplexing. Another area where considerable effort is being invested is the development of techniques for highly parallel resequencing. This will become useful for the discovery of new mutations.

The development of powerful techniques for multiplex or parallel analyses, along with software tools for assay design and data analysis, are invaluable contributions. Nevertheless, some obstacles must be overcome to establish cost-efficient tools for the wide variety of applications that are presently in demand, or that will be in the near future. Some of the areas where capability has been lacking are the highly multiplexed and directed amplification of genomic sequences, the design of complex oligonucleotide probes, and the cost-efficient synthesis of long oligonucleotides. The present work has been aimed at providing some of these capabilities.

The selector method, assisted by the PieceMaker software, is an ideal procedure for the preparation of samples for parallel resequencing of genes or exons. This will enable detailed analysis of large numbers of genomic regions in a sample in order to find the genetic causes of cancer and other diseases. The selector method will also serve as a probing method to establish the relative copy-number levels of specific sequences in different DNA samples. This allows moderately multiplexed analysis of copy number variations and of DNA methylation.

The focus of this work has been on providing computer support for various reagent design tasks. The aim has been to assist the search for combinations of sequences and sequence elements that function well together in a set of probes. For some types of probes, this task is solved by the ProbeMaker software. As new methods are constantly being developed, new applications are being found, and new readout techniques are invented,

the need for an even more flexible and versatile tool has become apparent. The Comodo system is to be regarded as a first approach to obtaining unified models for oligonucleotide design. Further development will be necessary for the system to fulfill its intended role. Ideally, this would be achieved by the combined effort of developers with expertise in different areas of reagent design and software development. Once complete, the system has the potential to significantly reduce the time required to develop new oligonucleotide design applications. This will remove one limiting factor in the development of new analysis methods, and it will free up often limited software development resources for other tasks.

Another challenge in oligonucleotide design is to define relevant design and selection criteria. One approach is to choose criteria that are more stringent than what is expected to be necessary, to ensure that the resulting probes will function well. The experimental testing of large sets of probes designed with different criteria could potentially increase assay success rate by defining new design rules, and it could also increase design success rate by allowing the tuning of criteria so that probe candidates are not unnecessarily rejected. The software described herein will be valuable also for this kind of studies.

A remaining challenge is that of cost-efficiently producing large numbers of long oligonucleotide probes as required in many methods, including the selector technique. Regular oligonucleotide synthesis of hundreds or thousands of probes is costly, while *in situ* synthesis methods for microarray production do not produce enough of each probe to last more than at most one or a few reactions. Pools of long oligonucleotides have proven useful for gene synthesis applications⁶³. Similarly, sufficient amounts of probes can be produced from small amounts of starting material by using RCA or C2CA (Howell *et al.*, in preparation). The ability to create large amounts of thousands of probes from an *in situ*-synthesized pool of starting material will significantly reduce the cost associated with many present and future methods.

Acknowledgements

The work presented here has been funded by Knut och Alice Wallenbergs stiftelse through the Wallenberg consortium north for functional genomics, and by the European Union through the sixth framework program integrated project MolTools.

I am grateful to many who have contributed in different ways to make this work possible.

First of all, I wish to thank my supervisor Mats Nilsson and co-supervisor Ulf Landegren. You have been unbounded sources of problems for me to solve, and you have given me the freedom to choose from these problems and to solve them in my own ways. Also, you have created an inspirational environment that I have enjoyed working in.

Many thanks also to all past and present members of the molecular diagnostics and molecular medicine groups. You are the ones who make up this environment which it has been interesting, exciting, and sometimes even fun to be a part of. In particular I wish to thank Fredrik Dahl for the fruitful collaboration on the selector project and Johan Banér and Jonas Jarvius for helping me to get started in the early days, Magnus Isaksson and Mathias Howell for good collaborations, and Lena Spångberg for excellent technical assistance.

This work has been conducted at the department of genetics and pathology at the Rudbeck laboratory. The staff, B-M, Ulla, and others, have performed an excellent job in minimizing the frictions often associated with the administrative side of things. All the other students, Ph.D. students, and faculty have contributed to making this place a fun place to be. A particular thanks for all the pub evenings and floorball games.

I also wish to thank all my former teachers who have somehow succeeded in keeping my interest for gaining new knowledge up during nearly 20 years of schooling. My family has also played an important role in this, for which I am thankful.

Many thanks to all members of the open-source community; your work is an endless source of inspiration for me, and a great help in my own work. Thanks to the SourceForge.net team for providing an excellent open source software development web site.

Most of all, thank you Elisebeth, for your love and devotion, for listening, and for having the patience to bear with me during times of frustration as well as excitement.

References

1. NCBI Genome biology web site. <http://www.ncbi.nlm.nih.gov/Genomes/> (2006-03-15)
2. NCBI dbSNP web site. <http://www.ncbi.nlm.nih.gov/SNP/> (2006-03-15)
3. Syvänen, A. C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*, **2**, 930-942.
4. Howell, W. M., Jobs, M., Gyllenstein, U. and Brookes, A. J. (1999) Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nat Biotechnol*, **17**, 87-88.
5. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. and Fodor, S. P. A. (1996) Accessing Genetic Information with High-Density DNA Arrays. *Science*, **274**, 610-614.
6. Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W. M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat Biotechnol*, **21**, 1233-1237.
7. Matsuzaki, H., Loi, H., Dong, S., Tsai, Y. Y., Fang, J., Law, J., Di, X., Liu, W. M., Yang, G., Liu, G. *et al.* (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*, **14**, 414-425.
8. Landegren, U., Kaiser, R., Sanders, J. and Hood, L. (1988) A ligase-mediated gene detection technique. *Science*, **241**, 1077-1080.
9. Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B. P. and Landegren, U. (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, **265**, 2085-2088.
10. Fan, J. B., Yeakley, J. M., Bibikova, M., Chudin, E., Wickham, E., Chen, J., Doucet, D., Rigault, P., Zhang, B., Shen, R. *et al.* (2004) A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome Res*, **14**, 878-885.
11. Pastinen, T., Raitio, M., Lindroos, K., Tainola, P., Peltonen, L. and Syvänen, A. C. (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res*, **10**, 1031-1042.
12. Cahill, P., Bakis, M., Hurley, J., Kamath, V., Nielsen, W., Weymouth, D., Dupuis, J., Doucette-Stamm, L. and Smith, D. R. (2003) Exo-proofreading, a versatile SNP scoring technology. *Genome Res*, **13**, 925-931.
13. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. and Chee, M. S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, **37**, 549-554.
14. Lyamichev, V., Mast, A. L., Hall, J. G., Prudent, J. R., Kaiser, M. W., Takova, T., Kwiatkowski, R. W., Sander, T. J., de Arruda, M., Arco, D. A. *et al.* (1999) Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat Biotechnol*, **17**, 292-296.
15. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, 5463-5467.

16. Fan, J. B., Chen, X., Halushka, M. K., Berno, A., Huang, X., Ryder, T., Lipshutz, R. J., Lockhart, D. J. and Chakravarti, A. (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res*, **10**, 853-860.
17. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, **242**, 84-89.
18. Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad, H., Ronaghi, M., Willis, T. D., Landegren, U. *et al.* (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, **21**, 673-678.
19. Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R. and Gunderson, K. L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat Methods*, **3**, 31-33.
20. Grossman, P. D., Bloch, W., Brinson, E., Chang, C. C., Eggerding, F. A., Fung, S., Iovannisci, D. M., Woo, S. and Winn-Deen, E. S. (1994) High-density multiplex detection of nucleic acid sequences: oligonucleotide ligation assay and sequence-coded separation. *Nucleic Acids Res*, **22**, 4527-4534.
21. Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res*, **30**, e57.
22. Banér, J., Isaksson, A., Waldenstrom, E., Jarvius, J., Landegren, U. and Nilsson, M. (2003) Parallel gene analysis with allele-specific padlock probes and tag microarrays. *Nucleic Acids Res*, **31**, e103.
23. Tost, J. and Gut, I. G. (2002) Genotyping single nucleotide polymorphisms by mass spectrometry. *Mass Spectrom Rev*, **21**, 388-418.
24. Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A. and Tunnacliffe, A. (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics*, **13**, 718-725.
25. Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J. *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A*, **99**, 5261-5266.
26. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350-1354.
27. Landegren, U. and Nilsson, M. (1997) Locked on target: strategies for future gene diagnostics. *Ann Med*, **29**, 585-590.
28. Broude, N. E., Zhang, L., Woodward, K., Englert, D. and Cantor, C. R. (2001) Multiplex allele-specific target amplification based on PCR suppression. *Proc Natl Acad Sci U S A*, **98**, 206-211.
29. Shaper, M. H., Zhang, J., Loraine, A., Liu, W., Di, X., Liu, G. and Jones, K. W. (2004) MARA: a novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays. *Nucleic Acids Res*, **32**, e181.
30. Callow, M. J., Drmanac, S. and Drmanac, R. (2004) Selective DNA amplification from complex genomes using universal double-sided adapters. *Nucleic Acids Res*, **32**, e21.
31. Banér, J., Nilsson, M., Mendel-Hartvig, M. and Landegren, U. (1998) Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res*, **26**, 5073-5078.

32. Lizardi, P. M., Huang, X., Zhu, Z., Bray-Ward, P., Thomas, D. C. and Ward, D. C. (1998) Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet*, **19**, 225-232.
33. Dahl, F., Banér, J., Gullberg, M., Mendel-Hartvig, M., Landegren, U. and Nilsson, M. (2004) Circle-to-circle amplification for precise and sensitive DNA analysis. *Proc Natl Acad Sci U S A*.
34. Yeakley, J. M., Fan, J. B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S. and Fu, X. D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol*, **20**, 353-358.
35. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. and Davis, R. W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet*, **14**, 450-456.
36. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**, 1675-1680.
37. Nygren, A. O., Ameziane, N., Duarte, H. M., Vijzelaar, R. N., Waisfisz, Q., Hess, C. J., Schouten, J. P. and Errami, A. (2005) Methylation-specific MLPA (MS-MLPA): simultaneous detection of CpG methylation and copy number changes of up to 40 sequences. *Nucleic Acids Res*, **33**, e128.
38. Banér, J., Marits, P., Nilsson, M., Winqvist, O. and Landegren, U. (2005) Analysis of T-Cell Receptor V β Gene Repertoires after Immune Stimulation and in Malignancy by Use of Padlock Probes and Microarrays. *Clin Chem*.
39. Szemes, M., Bonants, P., de Weerd, M., Banér, J., Landegren, U. and Schoen, C. D. (2005) Diagnostic application of padlock probes--multiplex detection of plant pathogens using universal microarrays. *Nucl. Acids Res.*, **33**, e70-.
40. Rychlik, W. and Rhoads, R. E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucleic Acids Res*, **17**, 8543-8551.
41. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132**, 365-386.
42. Kaderali, L., Deshpande, A., Nolan, J. P. and White, P. S. (2003) Primer-design for multiplexed genotyping. *Nucleic Acids Res*, **31**, 1796-1802.
43. Chen, S. H., Lin, C. Y., Cho, C. S., Lo, C. Z. and Hsiung, C. A. (2003) Primer Design Assistant (PDA): A web-based primer design tool. *Nucleic Acids Res*, **31**, 3751-3754.
44. Kampke, T., Kieninger, M. and Mecklenburg, M. (2001) Efficient primer design algorithms. *Bioinformatics*, **17**, 214-225.
45. Kaderali, L. and Schliep, A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340-1349.
46. Raymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G. and Fayard, J. M. (2004) ROSO: optimizing oligonucleotide probes for microarrays. *Bioinformatics*, **20**, 271-273.
47. Nordberg, E. K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365-1370.
48. Rouillard, J. M., Zuker, M. and Gulari, E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res*, **31**, 3057-3062.
49. Li, X., He, Z. and Zhou, J. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res*, **33**, 6114-6123.

50. Wetmur, J. G. (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Crit Rev Biochem Mol Biol*, **26**, 227-259.
51. von Ahsen, N., Wittwer, C. T. and Schutz, E. (2001) Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem*, **47**, 1956-1961.
52. Breslauer, K. J., Frank, R., Blocker, H. and Marky, L. A. (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*, **83**, 3746-3750.
53. Owczarzy, R., Vallone, P. M., Gallo, F. J., Paner, T. M., Lane, M. J. and Benight, A. S. (1997) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*, **44**, 217-239.
54. SantaLucia, J., Jr. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct*, **33**, 415-440.
55. Lyamichev, V., Brow, M. A. and Dahlberg, J. E. (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science*, **260**, 778-783.
56. Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. (1991) Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A*, **88**, 7276-7280.
57. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403-410.
58. Shendure, J., Mitra, R. D., Varma, C. and Church, G. M. (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet*, **5**, 335-344.
59. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.
60. Roberts, R. J., Vincze, T., Posfai, J. and Macelis, D. (2005) REBASE--restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*, **33**, D230-232.
61. Fredriksson, S., Gullberg, M., Jarvius, J., Olsson, C., Pietras, K., Gustafsdottir, S. M., Östman, A. and Landegren, U. (2002) Protein detection using proximity-dependent DNA ligation assays. *Nat Biotechnol*, **20**, 473-477.
62. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045-3054.
63. Richmond, K. E., Li, M. H., Rodesch, M. J., Patel, M., Lowe, A. M., Kim, C., Chu, L. L., Venkataramaiah, N., Flickinger, S. F., Kaysen, J. *et al.* (2004) Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res*, **32**, 5011-5018.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 148*

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-6832



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2006