Multi-node Graphs: A framework for Multiplexed Biological Assays

Noga Alon * Vera Asodi † Charles Cantor ‡ Simon Kasif § John Rachlin ¶

Abstract

Multiplex Polymerase Chain Reaction (PCR) is an extension of the standard PCR protocol in which primers for multiple DNA loci are pooled together within a single reaction tube, enabling simultaneous sequence amplification, thus reducing costs and saving time. Potential cost saving and throughput improvements directly depend on the level of multiplexing achieved. Designing reliable and highly multiplexed assays is challenging because primers that are pooled together in a single reaction tube may cross-hybridize, though this can be addressed either by modifying the choice of primers for one or more amplicons, or by altering the way in which DNA loci are partitioned into separate reaction tubes. In this paper, we introduce a new graph formalism called a *multi-node graph*, and describe its application to the analysis of multiplex PCR scalability. We show, using random multi-node graphs that the scalability of multiplex PCR is

*Departments of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel.

†Department of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel.

‡Department of Biomedical Engineering, Boston University, Boston MA, USA.

§Department of Biomedical Engineering, Boston University, Boston MA, USA, Center for Advanced Genomic Technology, Boston University, Boston MA, USA and Children's Hospital, Informatics Program, Harvard Medical School.

¶Department of Computer Science, Boston University, Boston MA, USA.

constrained by a phase transition, suggesting fundamental limits on efforts to improve the cost-effectiveness and throughput of standard multiplex PCR assays. In particular, we show that when the multiplexing level of the reaction tubes is roughly $\Theta(\log(sn))$ (where s is the number of primer pair candidates per locus and n is the number of loci to be amplified), then with very high probability we can 'cover' all loci with a valid assignment to one of the tubes in the assay. However, when the multiplexing level of the tube exceeds these bounds, there is no possible cover and moreover the size of the cover drops dramatically. Simulations using a simple greedy algorithm on real DNA data also confirm the presence of this phase transition. Our theoretical results suggest, however, that the resulting phase transition is a fundamental characteristic of the problem, implying intrinsic limits on the development of future assay design algorithms.

1 Introduction

The polymerase chain reaction (PCR) [1, 2] is a fundamental laboratory technique that has played a significant role in the growth of the biotechnology industry over the past 25 years. It is a crucial step in a vast array of forensic and genomic applications [3, 4]. Multiplex PCR (Figure 1) is an important variant of the standard PCR protocol in which multiple DNA loci are amplified in parallel within a single reaction tube [5]. Multiplex PCR has many applications including forensic analysis and paternity testing [2], whole-genome sequencing [3], broad-spectrum pathogen identification, and high-throughput pharmacogenomic studies that aim to elucidate the connection between genetic variability, drug response, and disease susceptibility [4].

In the standard (single-amplicon or single-plex) PCR protocol, two short oligonucleotide primers (typically 15-25 bases long) hybridize to opposite strands of a strand-separated (de-

natured) DNA molecule. A forward primer hybridizes to one strand at one end of the target locus while a reverse primer hybridizes to the opposite strand. In the presence of a DNA polymerase, bounded primers are extended one base at a time in opposite directions resulting in a double-stranded clone of the original DNA molecule across a finite range bounded by the hybridization locations of the original primers. The standard PCR protocol involves repeated cycles of DNA strand separation, primer hybridization, and primer extension, with each cycle yielding roughly double of the number of DNA sub-sequences, and resulting in millions of identical copies of a targeted sub-sequence of the original DNA molecule sufficient to enable down-stream biochemical analysis.

Parallel amplification of multiple DNA loci using multiplex PCR offers greater throughput and is more cost-efficient (due to the lower cost of reagents) but imposes additional constraints in order for the reaction to work reliably [5]. One of the most important factors affecting the reliability of a multiplex PCR assay is the tendency of individual primer pairs within a single reaction tube to cross-hybridize and form primer dimers rather than binding to their intended DNA target. The formation of primer dimers can significantly reduce product yield for one or more of the target amplicons. Thus, primer selection must carefully consider pairwise primer compatibility, a problem that increases quadratically with the level of multiplexing desired.

While multiplex PCR protocols have been widely used in large genome centers, industry, and small labs for a number of applications, no bounds were generally known on the expected coverage one can achieve with a given level of multiplexing. In order to investigate such limits, we introduce a novel graph theoretical construct called a multi-node graph. Each node in a multi-node graph is associated with a set of representatives and edges connecting two multi-nodes are either active or inactive depending on the representative currently assigned to each incident multi-node. Multi-node graphs are a natural model for the multiplex PCR assay design problem, where the multi-nodes correspond to DNA loci to be amplified, node

representatives correspond to candidate forward/reverse primer pairs for each locus, each of which is valid for amplifying the specific DNA locus, and edges connect pairs of nodes whose chosen primers are mutually pairwise compatible according to some set of cross-alignment screening procedures. By 'mutually compatible' we mean that they will not cross-hybridize if placed in the same reaction tube.

The problem of selecting primers and assigning them to a set of multiplex reaction tubes is shown to be equivalent to finding a disjoint clique cover in a multi-node graph. In the next section we explain some of the issues involved in the selection of primers for single-plex and multiplex PCR. We then formally define our notion of a multi-node graph and explain its correspondence to the problem of designing multiplex PCR assays. In subsequent sections we show that a theoretical analysis on random multi-node graphs implies intrinsic limits on the level of multiplexing that can be achieved and then show that such limits are consistent with simulations using a simple greedy heuristic on human DNA sequences. We also present an efficient multi-stage matching algorithm that identifies a multi-node graph clique cover (corresponding directly to a multiplex PCR assay design) in a random multi-node graph.

2 Selecting compatible primers for multiplex PCR

The central problem of multiplex PCR is the identification of pairwise compatible primers. Primer selection criteria for single-plex reactions typically consider primer length, melting temperature, GC content, and whether or not individual primers will form hairpin loops or homodimers. For multiplex PCR design, local sequence alignment (Smith-Waterman), or more advanced ΔG-based techniques are effective at filtering out candidates that are likely to cross-hybridize, but these methods are not 100% accurate. Some typical primer-selection criteria are provided in Table 1 in the Appendix. Other primer design tools such as Primer3 [6] offer an even greater range of user-specified criteria for identifying, ranking, and selecting

forward and reverse primers for each DNA locus to be amplified. The specific design criteria used in this paper are given in Table 1 in the Appendix.

To improve experimental robustness, primer selection criteria can be made increasingly stringent, though doing so may severely limit the number of primer candidates one can identify, and thus undermine any attempt to discover a valid multiplex PCR assay. By contrast, loose selection criteria may make it easier to find a highly-multiplexed assay design, though the design may not be robust, and fail experimentally under laboratory conditions.

In general, primer selection stringency determines the probability that two random pairs of primers for distinct DNA loci will be deemed to be mutually compatible, and we find that this probability, in turn, impacts the overall scalability of multiplex PCR. We have estimated this probability for both randomly generated and real DNA sequences. We find that two individual primers have about a 10-20% chance of being incompatible using standard alignment test thresholds. In order for two pairs of DNA loci to be amplified in the same tube, both the forward and reverse primers for each loci must be pairwise compatible. In other words, for loci A and B with corresponding forward/reverse primer pairs $(A_{forward}, A_{reverse})$ and $(B_{forward}, B_{reverse})$ respectively, we must explicitly verify compatibility between four primer pairs: $(A_{forward}, B_{forward})$, $(A_{forward}, B_{reverse})$, $(A_{reverse}, B_{forward})$, and $(A_{reverse}, B_{reverse})$, it being assumed that the two additional pairwise combinations $(A_{forward}, A_{reverse})$ and $(B_{forward}, B_{reverse})$ were already pre-screened as part of the process of identifying viable primer pair candidates for each individual locus. This result implies that if we randomly choose a particular forward/reverse primer pair for each DNA locus to be amplified, the probability that the two loci are multiplex compatible is approximately 0.40 to 0.60, a result we have confirmed on simulations using real DNA data for human SNP loci. However, the actual compatibility probability ultimately depends on various alignment threshold parameters used to identify putative primer-pair cross-interactions, and may be higher or lower depending upon the de-

sign requirements of a particular assay. This measure of compatibility defines the probability that two DNA loci can be successfully amplified together under laboratory conditions in a single reaction tube while avoiding the formation of primer-dimers that would otherwise interfere with the amplification process. Of course, the designer of a multiplex assay is free to choose a different primer pair for one or more loci, thus affecting which loci can be amplified together. The objective of multiplex PCR assay design thus involves the dual problem of selecting primers for each locus, and then partitioning the loci into disjoint multiplex-compatible sets, where the validity of a given partitioning scheme is directly impacted by the specific primers chosen for each target locus. The problem is naturally represented using a multi-node graph formalism that we now describe.

3 Multi-node graphs

Network and graph representations have been used with considerable success in computational biology as a mathematical foundation for wide-ranging of optimization problems [7, 8, 9, 10] including the analysis of multiplex PCR [11, 12, 13, 14]. The application addressed by this paper is concerned with multiplexing of PCR, the simultaneous amplification of multiple DNA loci from a single DNA sample. Multiplex PCR is a precursor to downstream biochemical analysis and readout. In SNP genotyping applications, the readout or 'calling' of SNPs is a separate multiplexing problem whose design parameters are impacted by the specific technology employed, and has been previously studied [15, 16].

In order to model the multiplex PCR problem we use a new framework of multi-node graphs. Each pair of PCR forward/reverse primers that can be used to amplify a DNA region is associated with a node in the graph. The entire set of forward/reverse candidate primers for a single specific DNA region is therefore represented by a set of nodes, called a multi-node. Thus each multi-node is uniquely associated with a particular DNA region (or SNP in

a genotyping assay). There are no internal edges among nodes in a multi-node. Two nodes are connected with an edge if the corresponding forward/reverse primer pairs associated with these nodes are mutually compatible. 'Mutually compatible' means that the primer pairs are unlikely to form primer-dimers due to their inherent sequence characteristics, thus enabling their corresponding loci to be amplified simultaneously as part of a single PCR reaction. Each tube in an assay will correspond to a set of nodes chosen from a graph, at most one node per multi-node (DNA region). Clearly, in order for the multiplex reaction to be effective all the nodes chosen to be included in a single tube must be mutually compatible i.e., form a clique in the graph. Thus, in order to amplify all regions, we have to choose one node from each multi-node, and cover the induced subgraph by cliques. Definition 3.1 below formalizes the concept of multi-node graphs.

Definition 3.1 *A multi-node graph is a graph $G = (V, E)$ with a partition of V into disjoint sets $V_1, \ldots, V_n$ called multi-nodes. Each multi-node is an independent set in G. An assignment A of representatives to multi-nodes is a function $A : \{V_1, \ldots, V_n\} \rightarrow V$ such that for all i, $A(V_i) \in V_i$. Let G_A denote the subgraph of G induced by $\{A(Vi) \mid 1 \leq i \leq n\}$.*

In problems related to multi-node graphs we are interested in subgraphs G_A that satisfy prescribed properties such as whether they contain cliques, independent sets, or other connectivity and topological characteristics. In Section 4 we study the clique cover problem for multi-node graphs.

An example multi-node graph having four multi-nodes is shown in Figure 2 along with three unique sub-graphs induced by a specific choice of representatives for each multi-node. Note that multi-nodes can contain varying numbers of representatives and that there is a many-to-one relationship between the representative assignment and the induced sub-graph.

Multi-node graphs are reminiscent of constraint networks [17, 18], and are closely related to problems such as Group TSP [19] and Group Steiner tree [20]. In addition, many important

graph problems for simple graphs have natural analogs for multi-node graphs including the clique cover problem.

The multi-node graph version of the disjoint clique cover problem asks whether there exists an assignment A of representatives to multi-nodes such that the resulting subgraph G_A can be covered by disjoint cliques of size k. The multi-node disjoint clique problem has direct application to the problem of designing multiplex PCR assays addressed in this paper. Specifically, we wish to identify the maximal value of k for which there exists with high probability a disjoint cover of all or most of the multi-nodes using k-cliques, in a random multi-node graph. For example, one of the subgraphs shown in Figure 2b enables the multi-node graph to be covered with two disjoint 2-cliques: $\{x, y\}$ and $\{u, v\}$. There is, however, no possible state-assignment inducing even a single 3-clique. We show that a multi-node graph formulation allows one to derive performance bounds that can guide the design of practical assays.

4 Results

4.1 A phase transition in achieving high-coverage / high-multiplexing assays

Given a multi-node graph G and an integer k, we would like to find an assignment A of representatives to multi-nodes such that we can cover G_A (or most of it) by pairwise vertex-disjoint cliques of size k. Our objective is to estimate the largest k for which such a cover exists in the random model. This provides a maximum efficiency solution subject to defined constraints. A solution that fully covers G_A corresponds to an assignment of every DNA locus (or rather their corresponding forward and reverse primers) to some reaction tube.

Let $G = (V, E)$ be a random multi-node graph with multi-nodes $V_1, \ldots, V_n$ of size s each,

where for all pairs of vertices u, v from different multi-nodes, we put an edge (u, v) in E independently with probability p. Let $N = |V| = ns$. We consider constant p and large N.

It is known (see, e.g. [21]) that the clique number of $G(N, p)$ is almost surely (that is, with probability that tends to 1 as n tends to infinity) $(2-o(1)) \log_{1/p} N$. Hence, for $k > 2 \log_{1/p} N$, there is almost surely no clique of size k in the random multi-node graph G, as it is a subgraph of the random graph $G(N, p)$ obtained by removing the edges inside the multi-nodes.

On the other hand, we show that if $k \leq (1 - o(1)) \log_{1/p} N$ and k divides n, then there is almost surely an assignment A of representatives to multi-nodes such that the graph G_A can be covered by pairwise vertex-disjoint cliques of size k. We further show that if $k \leq (2-o(1)) \log_{1/p} N$ then there is almost surely an assignment of representatives to multi-nodes such that $1 - o(1)$ of the vertices of G_A can be covered by pairwise vertex-disjoint cliques of size k.

Thus, when the size of the tube is roughly $2 \log(sn)$ (where s is the number of primer pair candidates per locus and n is the number of loci to be amplified), then with very high probability we can assign each and every amplicon to some multiplex tube. This partitioning of amplicons to tubes defines a disjoint clique cover on the corresponding multi-node graph. However, when the size of the tube exceeds these bounds, there is no possible cover and moreover the size of the cover drops dramatically, implying a phase transition on locus amplification coverage when the size of the tube is $2 \log(sn)$. In essence, designing full-coverage multiplex PCR assays is relatively easy for multiplexing levels just below a certain threshold, and difficult if not impossible just above that same threshold.

As described in subsection 4.2, we can confirm the presence of this phase transition using a simple greedy first-fit heuristic on real data. However, our theoretical results suggests that the resulting phase transition is a fundamental property of the problem itself, implying intrinsic bounds on the development of future assay design algorithms.

Proposition 4.1 *Let $G = (V, E)$ be a random multi-node graph with multi-nodes $V_1, \ldots, V_n$ of size s each, and with edge probability p. Let $N = |V| = ns$, and assume that $k \leq (1 - o(1)) \log_{1/p} N$ and that k divides n. Then there is almost surely an assignment A of representatives to multi-nodes such that the graph G_A can be covered by pairwise vertex-disjoint cliques of size k. Moreover, such an assignment can be found efficiently.*

Proof: We use a similar technique to the one used in [22]. Assume $n = mk$. Let $I_1, \ldots, I_k$ be an arbitrary partition of $[n] = \{1, \ldots, n\}$ into disjoint subsets of size m each. For an assignment A of representatives to multi-nodes, let $U_i = \{A(V_j) \mid j \in I_i\}$. We would like to find an assignment A for which the graph G_A contains m pairwise vertex-disjoint cliques of size k each, where each clique contains one vertex from every U_i. We now show that we can almost surely find such an assignment. We construct A by choosing the sets U_i one by one, so that for all $1 \leq i \leq k$, the subgraph induced by $U_1 \cup \ldots \cup U_i$ contains m pairwise vertex-disjoint cliques of size i each, where each clique contains one vertex from every U_j, $1 \leq j \leq i$. Let $W_1, \ldots, W_m$ denote the vertices of these cliques during the procedure. First choose U_1 arbitrarily and define $W_1, \ldots, W_m$ accordingly. Now suppose we have already found sets $U_1, \ldots, U_i$ with the required properties. In order to choose U_{i+1}, define a bipartite graph $H = (A, B, F)$, where $A = \{a_j \mid j \in I_{i+1}\}$, $B = \{b_1, \ldots, b_m\}$, and $(a_j, b_l) \in F$ if and only if there is a vertex $u \in V_j$ such that for all $w \in W_l$, $(u, w) \in E$. That is, $(a_j, b_l) \in F$ if and only if V_j contains a vertex that can be added to W_l to form a clique of size $i + 1$. If H contains a perfect matching $M \subseteq F$, then for all edges $(a_j, b_l) \in M$, define $A(V_j)$ to be a vertex $u \in V_j$ for which $(u, w) \in E$ for all $w \in W_l$ and add u to W_l. The set U_{i+1} obtained by this assignment has the required properties.

We now show that if $k \leq \log_{1/p} N - 3 \log_{1/p} \log_2 N$, n is sufficiently large and $s \leq n^{O(1)}$, then the probability that there is no perfect matching in H is at most $O(\frac{N^{-\Omega(\log_2 N)}}{k})$ in each stage. Therefore, the probability that there is a perfect matching in each stage, and thus there

is an assignment with the required properties, is at least $1 - O(N^{-\Omega(\log_2 N)})$.

For all $j \in I_{i+1}, 1 \leq l \leq m$,

$$Pr\left((a_j, b_l) \in F\right) = 1 - (1 - p^i)^s \geq 1 - (1 - p^k)^s.$$

Let $q = 1 - (1 - p^k)^s$. The events $(a_j, b_l) \in F$, for $a_j \in A, b_l \in B$, are independent. Thus, the graph H in each stage is a random bipartite graph with edge probability at least q. For $k \leq \log_{1/p} N - 3\log_{1/p}\log_2 N$

$$\begin{aligned} q &= 1 - (1 - p^k)^s \\ &\geq 1 - \left(1 - p^{\log_{1/p} N - 3\log_{1/p}\log_2 N}\right)^s \\ &= 1 - \left(1 - \frac{(\log_2 N)^3}{N}\right)^s \\ &\geq 1 - e^{-\frac{s(\log_2 N)^3}{N}} \\ &= 1 - e^{-\frac{(\log_2 N)^3}{n}}. \end{aligned}$$

Hence, by the known results on the existence of a perfect matching in random bipartite graphs (see, e.g. [23]), the probability that H contains no perfect matching is at most $O(me^{-mq})$. We show that this probability is $O(\frac{N^{-\frac{1}{2}\log_2 N}}{k})$.

$$\begin{aligned} me^{-mq} &\leq me^{-\frac{n}{k}(1 - e^{-\frac{(\log_2 N)^3}{n}})} \\ &\leq me^{-\frac{n}{\log_2 N - 3\log_2\log_2 N}(1 - e^{-\frac{(\log_2 N)^3}{n}})}. \end{aligned}$$

If n is sufficiently large and $s \leq n^{O(1)}$ then

$$1-\frac{(\log_2 N - 3\log_2\log_2 N)(\ln n + \frac{1}{2}\ln N \log_2 N)}{n} \geq e^{-2\frac{(\log_2 N - 3\log_2\log_2 N)(\ln n+\frac{1}{2}\ln N\log_2 N)}{n}} \geq e^{-\frac{(\log_2 N)^3}{n}}.$$

Therefore,

$$\begin{aligned} me^{-mq} &\leq me^{-\frac{n}{\log_2 N - 3\log_2\log_2 N}(1-e^{-\frac{(\log_2 N)^3}{n}})} \\ &\leq me^{-\ln n - \frac{1}{2}\ln N \log_2 N} \\ &= \frac{mN^{-\frac{1}{2}\log_2 N}}{n} \\ &= \frac{N^{-\frac{1}{2}\log_2 N}}{k}. \end{aligned}$$

Note that the above proof provides a polynomial time algorithm (see Figure 3) that almost surely finds the required assignment A and a partition of G_A into cliques of size k. □

Proposition 4.2 *Let $G = (V, E)$ be a random multi-node graph with multi-nodes $V_1, \ldots, V_n$ of size s each, and with edge probability p. Let $N = |V| = ns$ and assume $k \leq (2 - o(1))\log_{1/p} N$. Then there is almost surely an assignment A of representatives to multi-nodes such that $1-o(1)$ of the vertices of G_A can be covered by pairwise vertex-disjoint cliques of size k.*

Proof: Let μ denote the expected number of k-cliques in G. Then, $\mu = \binom{n}{k}s^k p^{\binom{k}{2}}$. For all $S \subseteq V$ with $|S| = k$ and $|S \cap V_i| \leq 1$ for all $1 \leq i \leq n$, let A_S be the event that S is a clique. For $S \neq T$, the events A_S and A_T are independent unless $|S \cap T| \geq 2$. Let $\Delta = \sum_{S,T} Pr(A_S \wedge A_T)$ where the sum is taken over all ordered pairs S, T such that $2 \leq |S \cap T| \leq k-1$. We now show that $\Delta \leq \frac{\mu^2 k^5}{2pN^2}$.

$$\Delta = \sum_{S,T} Pr(A_S \wedge A_T)$$

$$
\begin{aligned}
&= \sum_S Pr(A_S) \sum_T Pr(A_T \mid A_S) \\
&= \sum_S Pr(A_S) \sum_{i=2}^{k-1} \sum_{T:|S\cap T|=i} Pr(A_T \mid A_S) \\
&= \sum_S Pr(A_S) \sum_{i=2}^{k-1} \binom{k}{i}\binom{n-k}{k-i} s^{k-i} p^{\binom{k}{2}-\binom{i}{2}} \\
&= \mu \sum_{i=2}^{k-1} \binom{k}{i}\binom{n-k}{k-i} s^{k-i} p^{\binom{k}{2}-\binom{i}{2}} \\
&\leq \mu s^{k-2} p^{\binom{k}{2}} \sum_{i=2}^{k-1} \binom{k}{i}\binom{n-k}{k-i} p^{-\binom{i}{2}}
\end{aligned}
$$

As shown in [21], the largest term in the above sum is the first one. Thus,

$$
\begin{aligned}
\Delta &\leq \mu s^{k-2} p^{\binom{k}{2}} \sum_{i=2}^{k-1} \binom{k}{i}\binom{n-k}{k-i} p^{-\binom{i}{2}} \\
&\leq \mu s^{k-2} p^{\binom{k}{2}} (k-2)\binom{k}{2}\binom{n-k}{k-2}\frac{1}{p} \\
&= \mu s^{k-2} p^{\binom{k}{2}} (k-2)k(k-1)\binom{n}{k}\frac{(n-2k+3)(n-2k+4)\ldots n(k-1)k}{(n-k+1)(n-k+2)\ldots n}\cdot\frac{1}{2p} \\
&\leq \mu\binom{n}{k} s^k p^{\binom{k}{2}} \frac{k^5}{2ps^2n^2} \\
&= \frac{\mu^2 k^5}{2pN^2}
\end{aligned}
$$

We next show that the probability that G contains no clique of size k is $e^{-N^{2-o(1)}}$. Let Y be the maximal size of a family of pairwise edge-disjoint cliques of size k in G, and let $\mathcal{K}$ denote the family of all k-cliques of G. Then $\mu = E[|\mathcal{K}|] = \binom{n}{k} s^k p^{\binom{k}{2}}$. Let W denote the number of unordered pairs $\{S, T\}$ of k-cliques in G for which $2 \leq |S \cap T| < k$. Then $E[W] = \Delta/2$. Let $\mathcal{C}$ be a random subfamily of $\mathcal{K}$ chosen by putting each $K \in \mathcal{K}$ into $\mathcal{C}$ independently with probability $q = \frac{\mu}{\Delta}$. Let W' denote the number of unordered pairs $\{S, T\}$, $S, T \in \mathcal{C}$,

$2 \leq |S \cap T| < k$. Then $E[W'] = E[W]q^2 = \frac{\Delta q^2}{2}$. Delete from $\mathcal{C}$ one set from each such pair $\{S, T\}$, and denote the resulting family by $\mathcal{C}^*$. Then $\mathcal{C}^*$ is a family of pairwise edge-disjoint cliques of size k, and

$$\begin{aligned} E[Y] &\geq E[|\mathcal{C}^*|] \\ &\geq E[|\mathcal{C}|] - E[W'] \\ &= \mu q - \frac{\Delta q^2}{2} \\ &= \frac{\mu^2}{2\Delta} \\ &\geq \frac{pN^2}{k^5}. \end{aligned}$$

Let $Y_0, Y_1, \ldots, Y_m$, $m = \binom{n}{2}s^2$ be the edge exposure martingale on G, with the function Y as defined above. Y satisfies the Lipschitz condition, since adding or deleting an edge can change the number of pairwise edge-disjoint cliques by at most 1. By Azuma's inequality, the probability that G contains no clique of size k is

$$\begin{aligned} Pr(Y = 0) &= Pr(Y - E[Y] \leq -E[Y]) \\ &\leq e^{-\frac{(E[Y])^2}{2m}} \\ &\leq e^{-\frac{p^2N^4}{2\binom{n}{2}s^2k^{10}}} \\ &\leq e^{-\frac{p^2N^4}{N^2k^{10}}} \\ &= e^{-\frac{p^2N^2}{k^{10}}} \\ &= e^{-\frac{N^2}{O((\log_{1/p} N)^{10})}} \\ &= e^{-N^{2-o(1)}}. \end{aligned}$$

Set $m = n^{1-o(1)}$, and suppose $\binom{m}{k} s^k p^{\binom{k}{2}} > 1$ and $k = (2 - o(1)) \log_{1/p} ms = (2 - o(1)) \log_{1/p} N$. For any set of m multi-nodes, the subgraph of G induced by these multi-nodes contains a clique of size k with probability $1 - e^{-N^{2-o(1)}}$. There are $\binom{n}{m} < 2^n$ such sets. Hence, the probability that there is a set of m multi-nodes such that the subgraph of G induced by these multi-nodes does not contain a clique of size k is at most $2^n e^{-N^{2-o(1)}} = o(1)$. Thus we can almost surely find cliques of size k until there are less than $m = o(n)$ multi-nodes left. □

Note that this proof does not provide an efficient algorithm to find the desired assignment.

4.2 Confirmation of phase transition on human DNA data

In order to test whether our theoretical multiplex PCR performance predictions are consistent with real problems, we designed assays for $n = 20$ to $2,500$ SNPs randomly chosen from human chromosome 21, and catalogued in the dbSNP database [6]. All chosen SNPs were validated reference SNPs having at least 400 base pairs for high-complexity flanking sequence, necessary for identifying a reasonable number of candidate primer pairs. For each SNP, we generate a set of (forward, reverse) primer pair candidates using primer selection criteria provided in Table 1, resulting in 170.9 overlapping primer pair candidates per SNP locus, with 15.3% (386/2500) having no valid primer pairs. While our theoretical analysis assumes a fixed number of candidate primers per locus, this is clearly not the case in practice, as shown in Figure 4 which presents a distribution on the number of primer-pair candidates per locus using the primer selection criteria defined in Table 1 for the $n = 2,500$ selected SNPs. Our simulations show, however, that this variation in the number of primers per locus (corresponding to a variation in the number of representatives per multi-node) does not appear to impact the emergence of a theoretically predicted phase transition on coverage.

Multiplex PCR assays were then designed using a greedy *first-fit* assignment algorithm that

proceeded as follows: SNPs are processed in random order. For each SNP locus, we attempt to assign one of the primer pair candidates to a tube, with primers being systematically tested for tube compatibility in random order. A primer pair is compatible with a tube if both primers are compatible with all of the forward and reverse primers already assigned to the tube. If a given primer-pair is found to be incompatible with all tubes, we choose a different primer pair to test. If no primer pair is compatible with any tube, we pick a primer pair at random and assign it to a new tube. All SNPs, except those having no valid candidate primers, are thus assigned to some tube. Note that we have used a greedy approach rather than the matching algorithm described in Subsection 4.1 in order to save computation time. The experimental results suggest that the performence of this approach is similar.

Figure 5 is the result of a simulation using our greedy first fit algorithm. For varying numbers of SNP loci, we measured, over 10 random trials, the percentage of loci covered (i.e., successfully assigned to a 'full' k-plex tube) as a function of this multiplexing level, k. In each random trial, we attempted to assign SNP loci in a different random order. While each SNP locus had a fixed set of available primer candidates, the order in which different primer pairs were tested for compatibility also varied randomly from one trial to the next.

Figure 5 reveals the rapid reduction in locus coverage as the target multiplexing level is increased, consistent with our theoretical prediction of a phase transition on locus coverage. In addition, the figure shows how the onset of the predicted phase transition is slightly deferred as we increase substantially the number of loci to be amplified, n. In general, one expects coverage to increase with n because of the additional partitioning flexibility obtained. Surprisingly, however, the increases in coverage are not very substantial even with large increases in n, although coverage increases do depend on the specific level of multiplexing involved. Thus at the 10-plex level, a five-fold increase in the number of available loci, from $n = 500$ to $n = 2,500$ increases coverage from about 55% to almost 80% while at the 25-plex level, very

little impact on coverage is observed.

Further simulations suggest that the onset of an observed phase transition is critically dependent on various primer selection and cross-alignment parameters used to predict the potential formation of primer-dimers. One such parameter is the worst-case alignment (ΔG [kCal/mol]) occurring between the 3' tail of one primer anywhere along the sequence of another. A highly stringent ΔG of -4.0 kCal/mol induces an early phase transition, making the design of 20-plex assays more difficult. By contrast, a weaker stringency of -8.0 kCal/mol readily enables the design of 40-plex assays or higher for specific applications. While looser constraints enable one to design higher-multiplexing assays, they also introduce a greater probability that the assay will fail to yield some target amplicons under laboratory conditions. Tuning certain parameters of the actual PCR protocol (for example, modifying the concentration of specific primers) may enable one to overcome a limited number of amplification failures. Figure 6 suggests, however, that one cannot increase multiplexing without limit and that one eventually encounters a phase transition for any given level of primer selection and cross-alignment stringency. Naturally, there is a multitude of other constraints affecting the success rate of the assay design such as product sizes and competitive amplification that are not captured by the pairwise primer interaction model.

In general, we find that it requires an exponential number of DNA loci to be amplified to induce linear increases in the average level of multiplexing. This result is consistent with both our theoretical analysis and our greedy algorithm simulations, where we measured the average level of multiplexing as a function of the number of loci to be assayed (Figure 7). The figure provided shows average multiplexing for different multiplexing targets. For example, with a target of 10-plex (10 amplicons generated per tube), we limit tubes to no more than 10 primer pairs, deeming the tube to be full at that point. Primers must be assigned to some other tube, or to a new tube. The figure plots the average multiplexing level achieved for

varying numbers of DNA loci, up to $n = 2,500$ and for target multiplexing levels ($k = 10$, 15, and 20). We also allow for unlimited tube capacity ($k = \infty$). A log-linear best fit reveals that in the limit where tube size (multiplexing level) is unbounded, the average multiplexing level increases with the log of the number of DNA loci to be amplified ($R^2 = 0.98$).

5 Discussion

In this paper we introduced a new framework of multi-node graphs as a framework for multiplex PCR assay design, providing a basis for understanding the computational limits of multiplex scalability. We showed both theoretically and empirically that the attainable level of multiplexing increases in proportion to the log size of the locus pool. Additional empirical support for these bounds is described in [24, 25] where we experimented with several algorithms for assay design, some of which are also the basis for a freely-accessible web-enabled system for designing multiplex PCR assays [26] in a multi-objective decision-support context.

Beyond multiplex PCR assay design, multi-node graphs can also provide a convenient framework to study other important problems in computational biology such as context dependent biological networks and context specific biological interactions [24].

Computational biology has emerged as a rich area for new computational problems and formalisms. Multi-node graphs are one such new formalism, providing a useful representation for a number of problems in computational biology. Pooling experiments to increase parallelism subject to interaction constraints is a prolific application area and provides a useful way to parallelize many problems in biotechnology. Multi-node graphs have been shown as a rigorous framework for abstracting such problems, allowing computational biologists to derive both efficient solutions and fundamental limitations.

6 Acknowledgements

This work is supported in part by NSF grants DBI-0239435 and ITR-048715 and NHGRI grant #1R33HG002850-01A1. The authors thank Chumming Ding for collaboration on designing the software for multiplex PCR.

References

[1] Mullis, K., et al., Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol, 1986. 51 Pt 1: p. 263-73.

[2] Mullis, K.B., F. Ferr and R. Gibbs, The Polymerase chain reaction. 1994, Boston: Birkhäuser. xxii, 458.

[3] Latchman, D.S., PCR applications in pathology: principles and practice. Modern methods in pathology. 1995, Oxford; New York: Oxford University Press.

[4] Sninsky, J.J., M.A. Innis and D.H. Gelfand, PCR applications: protocols for functional genomics. 1999, San Diego: Academic Press. xviii, 566 p., [3] p. of plates.

[5] Edwards, M.C. and R.A. Gibbs, Multiplex PCR: advantages, development, and applications. PCR Methods Appl, 1994. 3(4): p. S65-75.

[6] Rozen, S. and H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol, 2000. 132: p. 365-86.

[7] Karp, R., Mapping the genome: some combinatorial problems arising in molecular biology. Proceedings of the twenty-fifth annual ACM symposium on the Theory of Computing, 1993: p. 278-285.

[8] Gusfield, D., Algorithms on strings, trees, and sequences: computer science and computational biology. 1997, Cambridge [England]; New York: Cambridge University Press. xviii, 534 p.

[9] Pevzner, P., Computational molecular biology: an algorithmic approach. 2000, Cambridge, Mass.: MIT Press. xviii, 314 p.

[10] Pevzner, P.A., H. Tang and M.S. Waterman, An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A, 2001. 98(17): p. 9748-53.

[11] Tettelin, H., et al., Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. Genomics, 1999. 62(3): p. 500-7.

[12] Doi, K. and H. Imai, A Greedy Algorithm for Minimizing the Number of Primers in Multiple PCR Experiments. Genome Inform Ser Workshop Genome Inform, 1999. 10: p. 73-82.

[13] Beigel, R., et al. An Optimal Multiplex PCR Protocol for Closing Gaps in Whole Genomes. in RECOMB. 2001.

[14] Kampke, T., M. Kieninger and M. Mecklenburg, Efficient primer design algorithms. Bioinformatics, 2001. 17(3): p. 214-25.

[15] Aumann, Y., E. Manisterski and Z. Yakhini, Designing Optimally Multiplexed SNP Genotyping Assays. Journal of Computer Systems Sciences, 2005. 70: p. 399-417.

[16] Sharan, R., J. Gramm, Z. Yakhini and A. Ben-Dor, Multiplexing Schemes for Generic SNP Genotyping Assays. Journal of Computational Biology, 2005. 12(5): 514-533.

[17] Dechter, R., Constraint processing. 2005.

[18] Kasif, S., Towards a Constraint-Based Engineering Framework for Algorithm Design and Application. 1996. 28(4es):66.

[19] Safra, S. and O. Schwartz, On the Complexity of Approximating TSP with Neighborhoods and Related Problems. Algorithms - ESA, Lecture Notes in Computer Science, 2003. 2832: p. 446 - 458.

[20] Garg, N., G. Konjevod and R. Ravi, A polylogarithmic approximation algorithm for the group Steiner tree problem. Proceeding of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

[21] S. Janson, T. Łuczak and A.Ruciński, Random Graphs, Wiley, 2000.

[22] N. Alon and Z. Füredi, Spanning subgraphs of random graphs, Graphs and Combinatorics 8 (1992), 91-94.

[23] B. Bollobás, Random Graphs, Academic Press, 1985.

[24] Rachlin J., D. Cohen, C. Cantor and S. Kasif, Biological Context Networks: A mosaic view of the interactome (in review).

[25] Rachlin, J., et al., Computational tradeoffs in multiplex PCR assay design for SNP genotyping. BMC Genomics, 2005. 6: p. 102.

[26] Rachlin, J., et al., MuPlex: multi-objective multiplex PCR assay design. Nucleic Acids Res, 2005. 33(Web Server issue): p. W544-7.

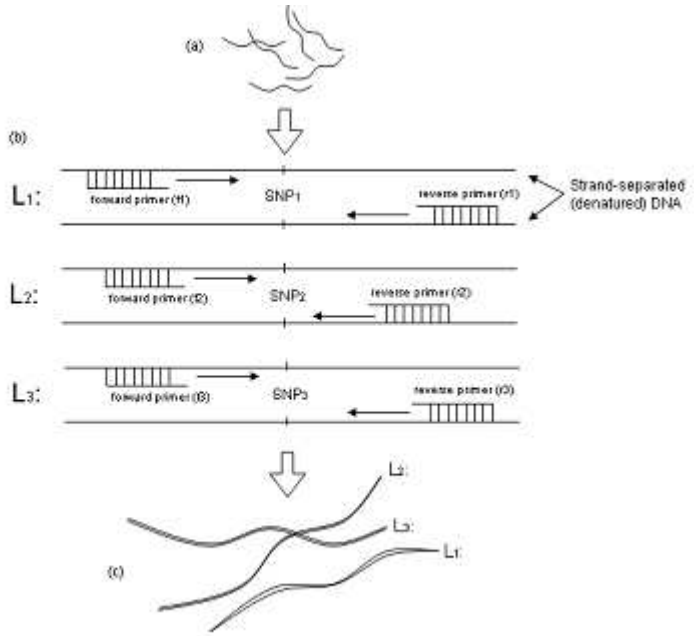

Figure 1: A schematic diagram of the multiplex PCR process. (a) Multiple primer pairs, each responsible for amplifying a single DNA locus, are pooled together in the same reaction tube. Individual primers must be designed to ensure that the formation of primer-dimers is avoided. (b) Forward and reverse primer pairs hybridize to a specific DNA locus following denaturation (strand separation) of the DNA molecule. In the presence of a DNA polymerase, primers are extended in opposite directions. (c) Reconstituted double-stranded amplicons produced in a single reaction tube, resulting in approximately double the amount of DNA after each repeated cycle.

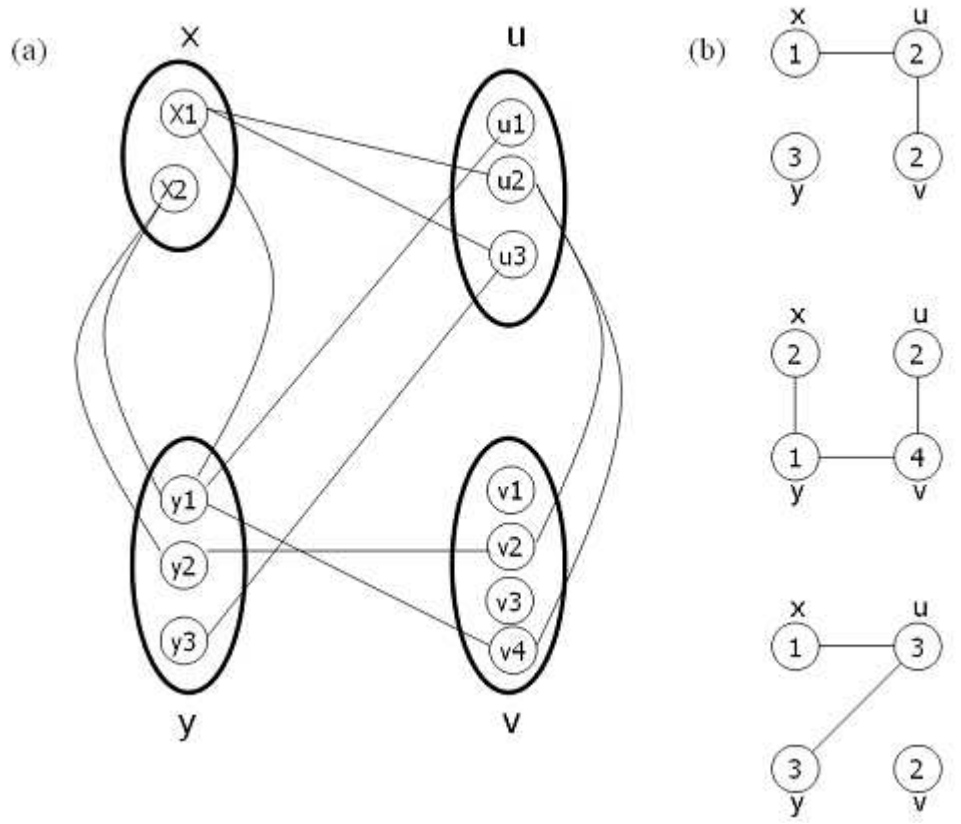

Figure 2: (a) The multi-node graph representation as a collection of independent sets. Each multi-node contains a unique collection of node states, only one of which is active at any given time. Multi-nodes may have varying numbers of states. The multi-node graph defines whether an edge is active between two multi-nodes as a function of the current state assignment. (b) The assignment of a specific state to each multi-node induces a particular graph topology, with three specific examples presented.

Matching Algorithm $(G(V = V_1 \cup \ldots \cup V_n, E), k)$

Partition $[n]$ into k disjoint sets of size $m = n/k$: $I_1, \ldots, I_k$
Let $I_1 = \{j_1, \ldots, j_m\}$
for $i \leftarrow 1$ to m **do**
  Choose $A(V_{j_i}) \in V_{j_i}$
  $W_i \leftarrow \{A(V_{j_i})\}$
end for

for $i \leftarrow 2$ to k **do**
  $A \leftarrow \{a_j \mid j \in I_i\}$
  $B \leftarrow \{b_1, \ldots, b_m\}$
  $F \leftarrow \{(a_j, b_l) \mid \exists u_{j,l} \in V_j \ \ \forall w \in W_l \ \ (u_{j,l}, w) \in E\}$
  Find a maximum matching M in $H = (A, B, F)$
  if M is not perfect **then**
    quit
  else
    for all $(a_j, b_l) \in M$ **do**
      $A(V_j) \leftarrow u_{j,l}$
      $W_l \leftarrow W_l \cup \{u_{j,l}\}$
    end for
  end if
end for

return $W_1, \ldots, W_m$

Figure 3: The matching algorithm for finding a clique cover of a multi-node graph.

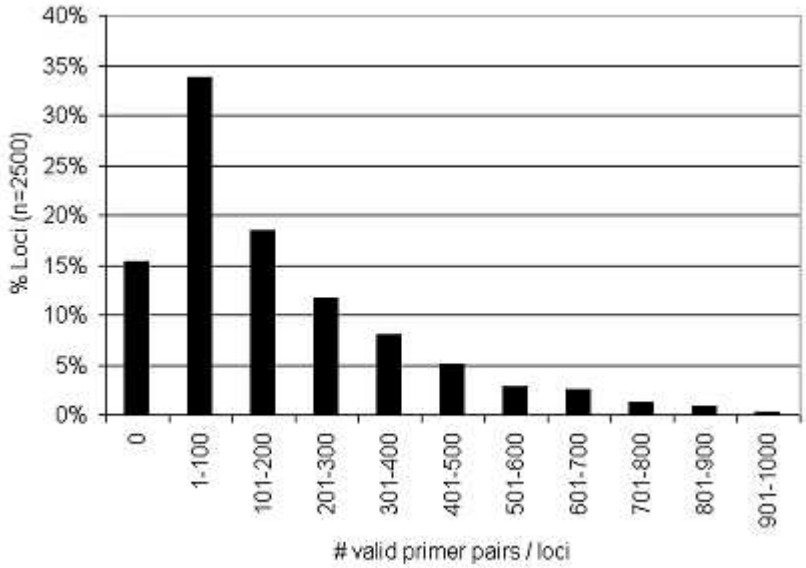

Figure 4: Number of primer pairs per SNP locus. The average number of primers per locus varies due to sequence variability. Approximately 15.3% (383/2500) had no viable candidate primer pairs. Avg. # candidates per loci = 170.9 (IQR=237).

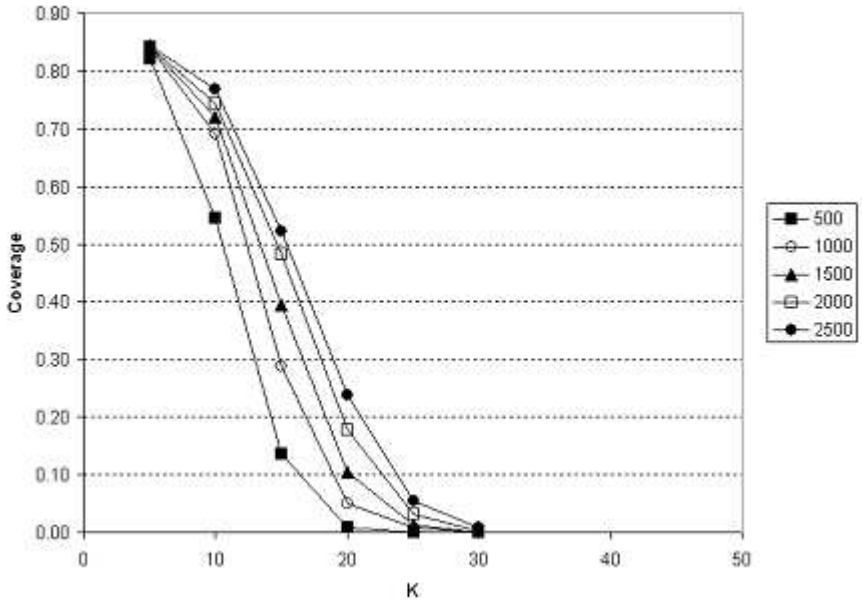

Figure 5: Locus coverage(% of genomic loci assigned to a k-plex tube) as the multiplexing level, k, is increased. Increases in k result in a corresponding decrease in the fraction of loci that can be amplified in k-plex tubes.

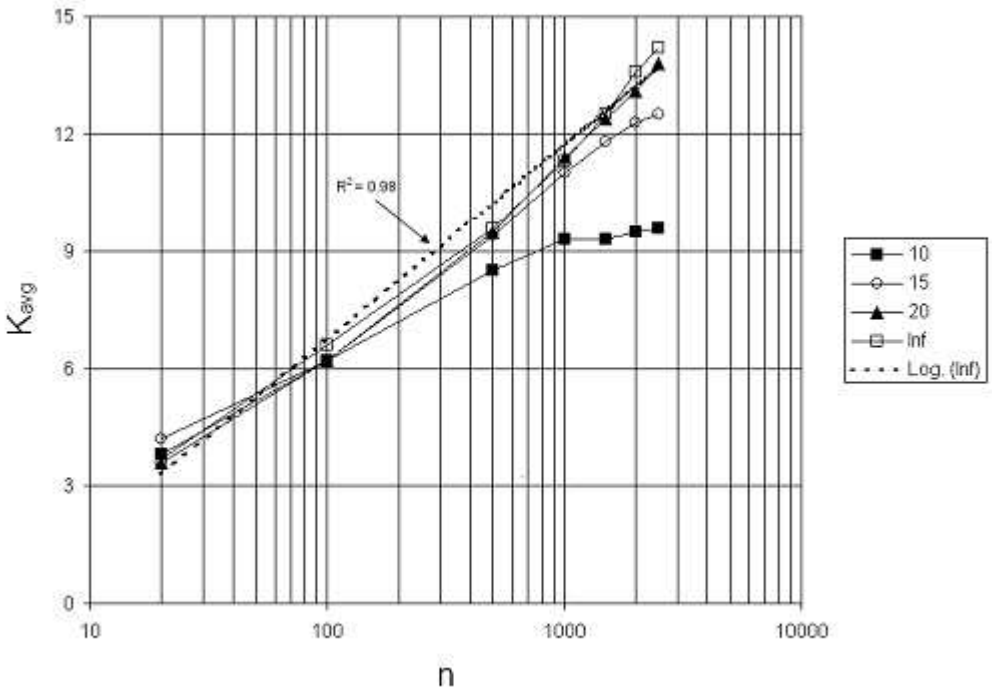

Figure 6: Average multiplexing level with increasing number of DNA loci to be amplified. More loci are added to the multiplex PCR assay design, average multiplexing level achieved increases in proportion to the log of the number of SNPs processed ($R^2 = 0.98$). Plot shown for varying bounds on the maximum multiplexing level per tube ($k = 10, 15, 20, \infty$).

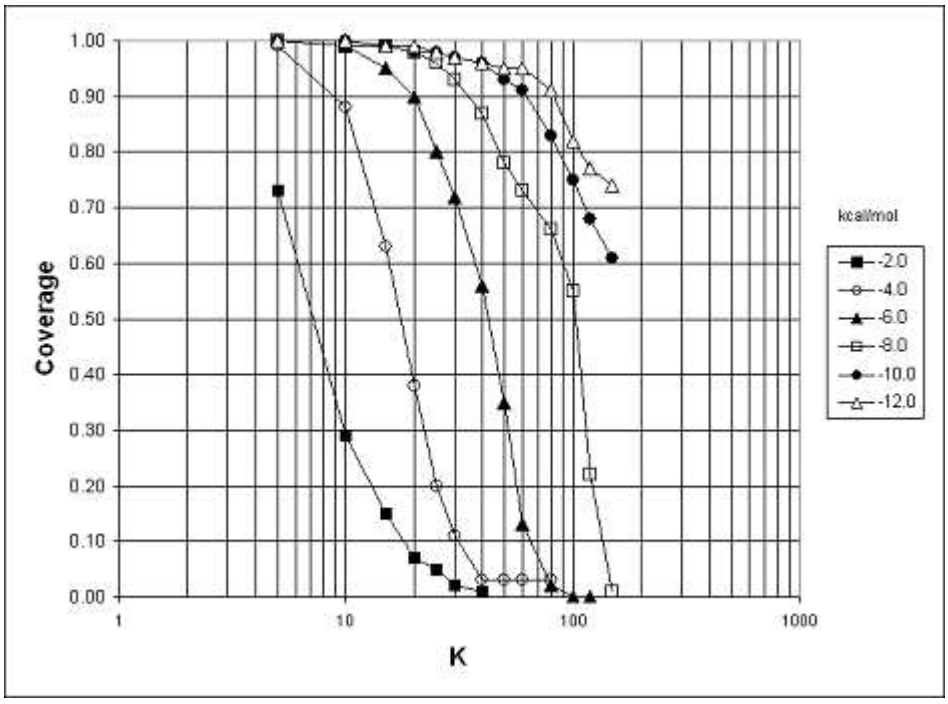

Figure 7: The onset of the phase transition from high to low coverage critically depends upon primer alignment thresholds used to detect the potential formation of primer dimers. Reducing the allowable worst case alignment between the 3' tail of one primer anywhere along the sequence of another primer within the same reaction tube (ΔG [kCal/mol]) from a highly stringent -4.0 kCal/mol to a relatively weak -8.0 kCal/mol enables one to achieve a substantially higher multiplexed assay design while maintaining high coverage, though a phase transition still occurs as the target multiplexing level (K) increases. (Simulations based on $n = 1200$ randomly selected chromosome 21 SNPs having at least one primer-pair candidate.)

Table 1: Primer design selection criteria. These criteria are used, where applicable, for determining the compatibility of forward and reverse primers within a given locus and for pairwise compatibility between primers for different loci. We employ the heuristic that if two sequences meet these primer selection criteria and share the same 3' position, only the shortest of the two primers is retained, since the shorter sequence is less likely to form primer-dimers.

Parameter	**Allowed Range**
Length	18-23 bases
% GC	30-70%
T_m (nearest neighbor)	57.0-63.0 C
T_m difference	3.0 C (for both candidate primer pairs and for all primers within a particular multiplex tube)
Base repeats	$\leq$ 3 bases maximum
Product Size	60-100 bases
Distance to SNP	150 bases (5' end) 5 bases (3' end)
Self complementarity local alignment score	$\leq$ 8.0 (match = 1.0, mismatch = -1.0, gap = -2.0)
3'-tail alignment ΔG	≥ -4.5 kCal/mol