How many Transcripts does it take to Reconstruct the Splice Graph?

Paul Jenkins[1], Rune Lyngsø[1], and Jotun Hein[1]

Dept. of Statistics, Oxford University, Oxford, OX1 3TG, United Kingdom;
{jenkins,lyngsoe,hein}@stats.ox.ac.uk

Abstract. Alternative splicing has emerged as an important biological process which increases the number of transcripts obtainable from a gene. Given a sample of transcripts, the alternative splicing graph (ASG) can be constructed—a mathematical object minimally explaining these transcripts. Most research has so far been devoted to the reconstruction of ASGs from a sample of transcripts, but little has been done on the confidence we can have in these ASGs providing the full picture of alternative splicing. We address this problem by proposing probabilistic models of transcript generation, under which growth of the inferred ASG is investigated. These models are used in novel methods to test the nature of the collection of real transcripts from which the ASG was derived, which we illustrate on example genes. Statistical comparisons of the proposed models were also performed, showing evidence for variation in the pattern of dependencies between donor and acceptor sites.

1 Introduction

Alternative splicing allows the creation of multiple mRNA transcripts from a single gene. Splicing takes place after the initial transcription of DNA into precursor (pre-) mRNA and before its translation. The process modifies pre-mRNA by discarding certain regions—known as *introns*—and retaining the rest. The resulting strand of ligated *exons*—retained sections—composes the mature mRNA, and by ligating different combinations of exons multiple mRNAs can be synthesised. Studies suggest that in many eukaryotes it is highly prevalent: as many as 74% of human genes undergo alternative splicing [1], with some genes able to produce a large number of different transcripts. Around 5% of human genes may each provide more than 100 putative transcripts [2]. Alternative splicing can therefore account for a number of otherwise unresolved problems, such as the discrepancy between the size of the human proteome and the smaller genome from which it is derived. It is also thought that alternative pre-mRNA splicing is a central mode of genetic regulation in higher eukaryotes (e.g. [3])—one well characterized example is the sexual identity of *Drosophila melanogaster* [4, 5]. Alternative splicing is therefore of central importance, and can now be studied in more depth thanks to the development of tools such as expressed sequence tags (ESTs) and, in recent years, microarray analyses [1, 6].

Exons can be spliced in different ways. Most exons are *constitutive*, that is, always retained in the mRNA. Exons either fully omitted or fully included are called *cassette*

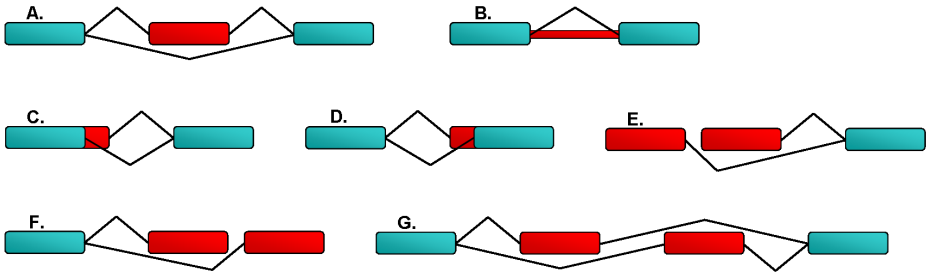

Fig. 1. Basic patterns of alternative splicing. Exons are shown as rectangles. Constitutive exons are in blue, regions which may be spliced out in red. Black lines represent paths of translation, from left to right. **A.** Cassette exon. **B.** Retained intron. **C.** Alternative 5' site. **D.** Alternative 3' splice site. **E.** Alternative promoter site. **F.** Alternative polyadenylation site. **G.** Mutually exclusive exons

exons. Alternative splice sites are also found within individual exons, known as alternative 5' or 3' splice sites. The mRNAs themselves may have alternative 5' or 3' ends, with alternative selection of the 5'-most or 3'-most exons. Finally a *retained intron* denotes an intron flanked by exons that is also included in the final mRNA. These 'building blocks' are illustrated in Fig. 1 (see also [4]). Some or all of these patterns may be observed from translations of a gene's mRNA, leading to potentially complex overall splicing patterns (Fig. 2).

Traditionally the transcriptome of a gene has been represented by an exhaustive list of its splice variants. However, as the prevalence of alternative splicing has become apparent, a need for more concise notation has emerged. Heber *et al*. [7] introduce the idea of the *alternative splice graph* (ASG), which enables the set of possible transcripts to be represented in a single graph, avoiding the error-prone nature of case-by-case transcript reconstruction. Denote the ASG as $G = (V, E)$, defined as follows. Let $\{s_1, ..., s_n\}$ be the set of RNA transcripts of the gene of interest, with each s_k corresponding to a sequence of genomic positions V_k ($V_i \neq V_j$ for $i \neq j$). Define $V := \bigcup_i V_i$, the set of all transcribed positions, and $E := \{(v, w) : v \text{ and } w \text{ form consecutive positions in at least one transcript } s_i\}$. Hence the ASG G is a directed graph and a putative transcript is any path in G. The graph is also acyclic, since the exons present in any spliced transcript are retained in the correct 5' to 3' linear order [8, 9]. Finally strings of consecutive vertices with $indegree = outdegree = 1$ are collapsed into a single vertex. So each exon fragment (i.e. portion of an exon bounded by two splice sites) is represented as a single vertex. This enables the ASG to be illustrated in a similar manner to that shown in Fig. 2—the numbered blocks are vertices, and the arcs read from left to right are directed edges. The ASG is a convenient, compact representation of all the splicing events associated with a particular gene, and lends itself to much further investigation.

Note the number of putative transcripts of an ASG equals or exceeds the number of distinct transcripts used to construct the ASG. The assumption that any path through the ASG represents a putative transcript in effect assumes that splicing events are independent. In this paper we propose two ASG based Markovian models of isoform generation

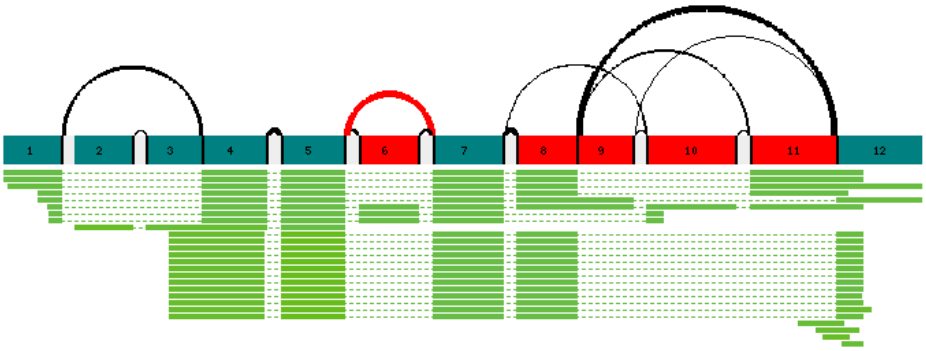

Fig. 2. An example of more complicated splicing patterns: human gene neurexin III-β (Ensembl ID ENSG00000021645, gene not to scale). Splicing events are represented by curved edges. The fragment labelled 6 is a cassette exon. More complicated and nested relationships are also visible. Edge thicknesses are proportional to EST support for that splicing event. ESTs from which the ASG was reconstructed are aligned below. Image derived from the Alternative Splicing Gallery [2]

to investigate this independence assumption. We further introduce simulation based and graph theoretical algorithms to investigate the question of whether the existing transcripts associated with a given gene are likely to have come from a random sample or from a strongly pruned subset of the transcriptome, either through non-independence of exons or through other effects such as ascertainment bias.

2 Transcript Generation Models

We propose two simple models of transcript generation, each utilising different parameter spaces. The process in our models is Markov in the sense that if we reach a particular exon fragment, the following fragment to be included does not depend on any other earlier decisions upstream. There exists a probability distribution over the transcriptome of a given gene, which can be modified conditional on additional knowledge, such as a cell's tissue type. For now we will not assume such further knowledge, which in many cases this will not unduly affect the distribution of interest. Tissue-specific control appears to be restricted to a relatively small number of specialised genes: only 2.2% of alternative splicing relationships have been observed with high confidence to be tissue-specific [10]. Tissue-specific control would cause a higher degree of exon coupling, since transcripts are effectively generated from two overlapping, yet distinct, sub-ASGs.

2.1 Model 1: Pairwise model

We approximate the discrete structure of a gene by an interval of the real line $[0, L]$. Superimposed on the gene is a (fixed) set V of exon fragments, a collection of subsets

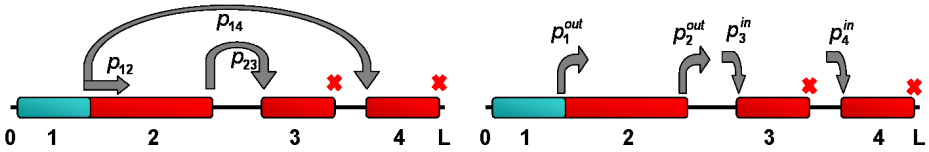

Fig. 3. Transcript generation of a simple example gene under each model. Constitutive exon shown in blue. Exons which may be spliced out shown in red. In this example label members of V as $1, \ldots, 4$, so that $I = \{1\}$, $p_1^{\text{start}} = 1$, $T = \{3, 4\}$; the read is terminated on reaching the 3' end of exon 3 or 4. (*Left*) Pairwise model. P is the zero matrix other than p_{12}, p_{23} and p_{14} ($p_{23} = 1$). (*Right*) In-out model. Here, $\mathbf{p}^{\text{in}} = (0, 0, p_3^{\text{in}}, p_4^{\text{in}})$ and $\mathbf{p}^{\text{out}} = (p_1^{\text{out}}, p_2^{\text{out}}, 0, 0)$ (with $p_2^{\text{out}} = 1$)

of the line (as in Fig. 2). Based on the underlying ASG there exist pairwise probabilities for each pair of fragments (v_1, v_2) that have been observed to be adjacent in at least one transcript, representing the probability that, as we read through the mRNA's sequence, if it contains v_1 we then jump forward to v_2. These probabilities can be captured as a $|V| \times |V|$ strictly upper triangular probability matrix P, with entries defined by p_{ij} = probability of a transcript jumping from fragment i to fragment j, given that it contains i.

To account for the features of Fig. 1, define sets $I, T \subseteq V$ of initiation fragments and terminal fragments, respectively. For each walk, rather than beginning at 0 proceed randomly with probability p_i^{start} from $i \in I$. Similarly, for each walk that reaches a fragment $t \in T$, transcript generation is terminated at the 3' end of that fragment. Thus, the complete model is captured by the collection $(V, P, I, \mathbf{p}^{\text{start}}, T)$ (Fig. 3). Intuitively this is the most general model consistent with observed splicing events that assumes independence between splicing events.

2.2 Model 2: In-out model

The pairwise model allows the modelling of dependencies between the donor and acceptor sites in a splicing event. As an alternative, we will also consider the most general model consistent with observed splicing events that models 'donation' and 'acceptance' of the splicing event independently. With each exon fragment $x \in V$, associate two probabilities p_x^{in}, p_x^{out}, the probabilities of jumping 'into' and 'out of' the the gene. Conceptually we can imagine travelling along the real line from 0 to L, and as we reach each exon fragment jumping 'in' with probability p^{in} if we are 'out', then jumping 'out' with probability p^{out} if we are 'in'. This in effect models inclusion of isolated exons as independent events, where each exon is included with a probability reflecting the strength of its acceptor site. Note that at most two probabilities are used at each fragment rather than up to $n = |V|$ for each in the pairwise model. The in-out model seeks to explain the splicing events we observe with only $O(n)$ parameters, compared to the $O(n^2)$ parameters of the pairwise model. $I, T \subseteq V$ and $\mathbf{p}^{\text{start}}$ are defined as in the pairwise model. Thus, the complete model is captured by the collection $(V, \mathbf{p}^{\text{in}}, \mathbf{p}^{\text{out}}, I, \mathbf{p}^{\text{start}}, T)$ (Fig. 3).

2.3 Hypothesis Testing

The in-out model is nested in the pairwise model; we can represent an in-out model $(S, \mathbf{p}^{\text{in}}, \mathbf{p}^{\text{out}}, I, \mathbf{p}^{\text{start}}, T)$ as a pairwise model $(S, P, I, \mathbf{p}^{\text{start}}, T)$ with $p_{ij} = p_i^{\text{out}} p_j^{\text{in}} \prod_{i<k<j}(1-p_k^{\text{in}})$. In a similar way, the pairwise model can be emdedded in what we'll refer to as model 0, that which simply assigns a probability to each putative transcript. Given a gene we can propose the following test for assessing the relative applicability of two models a, b, with $b \subseteq a$. For a given sample of transcripts $\{s_1, \ldots, s_n\}$ define the likelihood ratio statistic $\Lambda = \sup_{\mathbf{q_b}} L_b(\mathbf{q_b}) / \sup_{\mathbf{q_a}} L_a(\mathbf{q_a})$, where $\mathbf{q_a}, \mathbf{q_b}$ are the parameters under each model and L_a, L_b are the likelihoods of the data under each model, assuming independent sampling. The probability of each transcript is the product of the relevant probabilities involved in its generation. For example in Fig. 3, $P(1 \cup 2 \cup 3) = p_{12}p_{23}$ under the pairwise model and $P(1 \cup 2 \cup 3) = (1 - p_1^{\text{out}})p_2^{\text{out}}p_3^{\text{in}}$ under the in-out model. If the in-out model holds then $-2 \ln \Lambda \dot{\sim} \chi^2(z)$, where z is the number of degrees of freedom.

3 ASG Recovery Tests

3.1 Model Based Tests

Once we have a model describing transcript generation from an ASG, we can address the highly relevant question of the confidence we have in knowing the full true ASG. We propose a bootstrap-like method to assess the likelihood that the full ASG has been reconstructed, or alternatively to detect ascertainment biases in existing transcript databases, using a transcript generation model as follows. Assume that we have reconstructed an ASG from m transcripts. We may then ask what the probability is of drawing m independent samples from the full ASG that covers all edges in the full ASG. This can be computed exactly, albeit very inefficiently. Alternatively one can repeatedly sample m transcripts from the full ASG and check whether all edges are represented in these transcripts (or, if the in-out model is assumed, whether all choices are represented) to obtain a p-value for the scenario of recovering the full ASG from m transcripts.

Unfortunately, we do not necessarily know the full ASG but only the inferred ASG. So what can we expect if we sample from the inferred ASG? Assume that the inferred ASG is in fact the full ASG, and that the chosen model of transcript generation holds. Then we are indeed sampling from the full ASG and we can expect the rejection rate—i.e. the false negative rate—to equal one minus the p-value computed. If the inferred ASG does not coincide with the full ASG, the acceptance rate—i.e. the false positive rate—cannot be similarly tied to the p-value computed. Indeed if $m = 1$, the inferred ASG will offer only one putative transcript and our sampling test will always accept the inferred ASG. However, as shown in Section 4, the false positive rate does seem to follow the p-value threshold for realistic data. Intuitively, if after m transcripts the ASG is fully recovered, or close to it, then there is a higher probability of some redundancy in the real collection of transcripts—indicating that they do indeed cover the whole ASG.

Algorithm 1 Minimum Path Cover

while there is a non-cyclic path π from s to t in G_w **do**
  for all edges $e \in \pi$ **do**
    if $e \in E$ **then**
      $w(e) \leftarrow w(e) - 1$
    else
      $w(e) \leftarrow w(e) + 1$
  Recompute G_w

Alternatively, if in general sampling m transcripts does not recover the ASG then there is little redundancy in the collection, and hence a higher probability that there exist other undiscovered edges.

If testing whether a fraction α of the full ASG has been recovered, we are on even less solid ground. Sampling from the inferred ASG and accepting if a fraction α of the inferred ASG has been recovered, not even the false negative rate can be theoretically linked to the p-value computed. Assume that the full ASG offers three possible transcripts and that $m = 2$ and $\alpha = \frac{2}{3}$. With probability $\frac{2}{3}$ the inferred ASG will be based on two different transcripts, i.e. offer two possible transcripts. However, sampling from the inferred ASG we only achieve a p-value of $\frac{1}{2}$ for having recovered a fraction of α of the full ASG. Again we refer to Section 4 for empirical results on the usefulness of our computed p-value on realistic data.

3.2 ASG Based Tests

Without an accepted model for the alternative splicing observed for a gene, we cannot simulate transcript generation. We may however still make a qualitative assessment of the validity of the reconstructed ASG—or alternatively of whether transcripts are fully determined by regulatory factors rather than generated according to the combinatorial model implicit in the ASG representation—in the context of the transcripts used to reconstruct it by considering *informative* transcripts. A transcript is considered informative if it reveals one or more new edges of the ASG. A transcript corresponds to a path through the ASG. So a set of transcripts elucidating the full potential of the ASG uniquely corresponds to a set of paths covering all the edges in the ASG (i.e. a set of paths $\mathbf{P} = \{P_1, \ldots, P_k\}$ such that every edge of the ASG occurs in at least one path P_i in $\mathbf{P}$). For convenience we will assume that all paths have to start at source s and terminate at sink t. This can be realised by amending the ASG with s that has edges to all initiation fragments and t that all terminal fragments have an edge to. If $G = (V, E)$ denotes the ASG, it is a straightforward observation that the maximum number of informative transcripts is

$$2 + |E| - |V| \ . \tag{1}$$

The minimum number of informative transcripts is equivalent to a minimum path cover, a classic problem related to maximum flow (see e.g. [11]). For reference, algorithm 1

provides a simple augmenting path solution for reducing any path cover to a minimum path cover in time $O((|V| + |E|)\,|\mathbf{P}|)$ where $\mathbf{P}$ is the initial path cover. For each edge $e \in E$, its weight $w(e)$ is initialised to the number of paths covering e in the initial cover. Define $G_w = (V, E_w)$ where $E_w = \{e \in E \mid w(e) > 1\} \cup \{(v, u) \mid (u, v) \in E\}$. I.e. G_w contains all edges covered by more than one path, and the reverse edge of all the edges in G. At termination the minimum path cover size can be determined as the sum of the weights of the edges leaving the source node.

4 Results

We are interested in choosing a model relevant to an ASG constructed from real transcripts. Ideal for obtaining large-scale data on alternative splicing events is microarray technology, but this is still in its infancy, with only a handful of large-scale investigations into exon skipping events [12, 13]. Ultimately it is hoped that the ability to attach accurate inclusion rates to individual exons, and even the possibility of sampling full-length mRNA transcripts [14] will be possible. For illustrative purposes we must now content ourselves with using ESTs, whilst being mindful of their limitations [15], e.g. ESTs exhibit a strong bias for the 3' end of the gene. The Alternative Splicing Gallery [2] catalogues EST support for each human gene, from which maximum likelihood estimates (MLEs) for the probabilities associated with each exon fragment can be calculated via a simple transcript counting argument. We apply this to an example gene, Neurexin III-β; alternative splicing in neurexins has been well-characterized [16]. Consider Fig. 2. EST support for this gene suggests several distant exon coupling relationships, for example between exons 6 and 10. For convenience extend any partial EST to its full-length counterpart if this can be achieved unambiguously, otherwise omit it. A hypothesis test comparing model 0 against the pairwise model yields a p-value of 0.0026, confirming our suspicion that entirely independent splicing of exons may not be applicable for this gene.

For genes with larger ASGs, the cardinality of the set of all putative transcripts and hence the number of parameters required for use with model 0 can grow exponentially with the number of alternative splice sites, so that a large number of observations are required to accept model 0. At present these are generally lacking (suggesting that in fact the true ASG has not yet been observed—see Section 3.1), so for these genes we must either focus on short alternatively splicing regions, or instead we can test the relative merits of the pairwise and in-out models to provide some measure of the dependence in splicing between different exons. As an example consider the gene ABCB5, one of the 89 human genes known to offer more than 5000 putative transcripts [2]. It is a gene of interest also due to its association with drug resistance in human malignant melanoma, with both functional and non-functional splicing variants [17]. The likelihood ratio test was applied to four regions of the gene observed to exhibit alternative splicing. We make the additional assumption that these regions are bounded by constitutive exons, prohibiting under the models the splicing together of fragments from disparate regions of the gene (which would unnecessarily increase the parameter space in order to accommodate splicing events of negligible probability). p-values for the four regions are

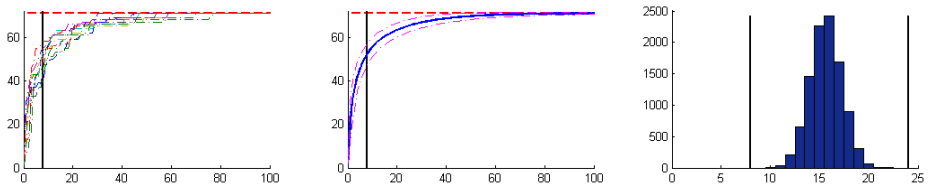

Fig. 4. (*Left*) Ten simulated reconstructions of the ASG for human gene ABCB5, under the pairwise model. Number of sampled transcripts (x-axis) is plotted against size of the reconstructed ASG (y-axis). Full ASG size shown as a dashed line. Minimal possible number of transcripts annotated as a vertical line. (*Centre*) Mean number of reconstructed edges across 10000 simulations ± 1 standard deviation. (*Right*) Histogram across 10000 simulations of number of informative transcripts. Maximum and minimum number of such transcripts are annotated

0.000, 0.029, 0.001, 0.000; the overall p-value is 0.000. All 89 genes were similarly tested: of them, 13 were deemed not to comprise any testable regions. Of the remaining 76 genes, 20 (26%) were accepted at the 5% level to be described by the in-out model. These seem to be the genes for which the assumption of independence between exons is most applicable.

We infer that ABCB5 is most suitably described by the pairwise model. Let us suppose then that transcripts are generated for ABCB5 under the pairwise model. Reconstruction of the ASG under this model is summarized in Fig. 4, with the minimal number of transcripts required to recover the ASG annotated. The size of the ASG is measured in the number of its recovered edges. The probabilities for the pairwise model are chosen using MLEs described previously. Consider Fig. 4(*left*). The 10 example simulations generally follow the growth curves one would expect of sampling with replacement. In some simulations the last few edges persist in remaining undiscovered even after the generation of 100 transcripts, but by 20 transcripts the mean proportion of the ASG to have been recovered is 90.8% (Fig. 4(*centre*)). What does this indicate about the probability that the 20 ESTs used to construct the ASG in the first place did in fact construct the complete ASG? If we apply our bootstrap-like method, none of the set of simulated transcript samples successfully recovers the full ASG resulting in a p-value of 0.

But how much can the p-value be trusted? To answer this we set up an experiment using the ABCB5 pairwise model as the true source for generating transcripts. From this we repeatedly sampled m transcripts and computed the p-value for the ASG inferred from these m transcripts. This was repeated for various choices of m. The outcome of this experiment is illustrated in Fig. 5(*top*). Both the fraction of graphs inferred from m transcripts coinciding with the full ASG, and the fraction of inferred graphs accepted at various acceptance rates are plotted. Encouragingly, it is evident from the righthand graph that there is a strong correlation between when we start to recover the full ASG and when we start to accept the inferred ASG. This indicates that our p-value does indeed capture whether the transcripts contain sufficient redundancy.

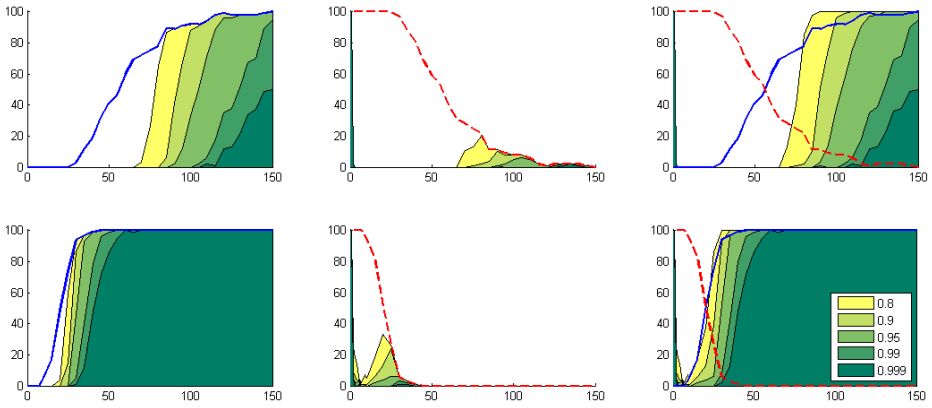

Fig. 5. Results of experiment described in text. Number of sampled transcripts (x-axis) is plotted against percentage of experiments (y-axis): percentage recovering the full ASG (*left*), percentage not recovering the full ASG (*centre*) and both (*right*). The fraction of such experiments for which the inferred ASG is accepted as the true ASG is shown for various confidence levels. The first row illustrates results for full recovery of the ASG, the second row for $\alpha = 90\%$ recovery of the full ASG

Note that the central graph, which plots acceptances of non-fully recovered inferred ASGs, separates the type II errors; any accepted graphs here are false positives. Similarly the lefthand graph, which plots acceptances of fully recovered inferred ASGs, separates the type I errors; any graph not accepted here is a false negative. As anticipated in Section 3.1, for very low m most experiments yield a high false positive rate, but in all our simulations this effect quickly dies away by $m = 3$.

For ABCB5, no acceptances are observed at the 20 transcript level, and we safely deduce that a scenario of independent random samples from a fully recovered ASG is not supported. This implies that either the ASG derived from the real 20 transcripts is a proper subset of the true ASG, or that the collection of transcripts used to infer the ASG is likely to be biased in the sense that there is little redundancy in the collection, and an emphasis rather on novel transcripts. As mentioned above, such an observation seems likely in a database with both human and biological biases. We performed a similar detailed investigation into all 56 genes with more than 5000 putative transcripts satisfying the pairwise model (data not shown) and found that in no cases was the reconstructed ASG accepted at confidence 0.95. Thus all could reasonably be said to exhibit a bias in their transcript records. This should of course be taken with the caveats associated with ESTs and the assumption that transcript generation is assumed to be correctly described by the pairwise model, along with the fact that by choosing complex genes to begin with these results will not be indicative of the rest of the genome. But this illustration offers a novel first step towards a method for teasing out the complex relationships discussed, which are not discernible from the ASG alone.

As mentioned in Section 3.1 we cannot expect our assessment of partial ASG recovery to be as precise as our assessment of full ASG recovery. To further investigate depen-

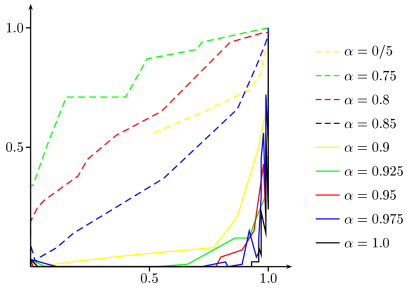

Fig. 6. Partial recovery results for nine different values of α, ranging from full ASG recovery to recovery of half the ASG. Fraction accepted as recovered to degree α at confidence level 0.95 is plotted against fraction recovered to degree α

dence on α of the quality of the p-value computed we ran experiments similar to those plotted in Fig. 5 for a range of α values with confidence level 0.95. Fig. 6 plots the fraction of accepted ASGs against the fraction of inferred ASGs containing at least a fraction α of the edges in the ABCB5 ASG. Ideally we would expect a phase transition from no accepted ASGs to all ASGs being accepted around the point where 95% of the inferred ASGs contain at least α of all edges. This is indeed observed for high values of α, but for α values lower than 0.9 there is an increasing tendency toward a mere linear relationship between ASG recovery and ASG acceptance. Remembering that ASG acceptance is more likely for a false positive than for a true positive it is thus clear that our method should not be applied for low α values.

5 Discussion

In this work we have proposed a mathematical framework to consider how to predict the nature of transcript generation in alternatively splicing genes. These models can be used make inferences on questions such as the levels of independence in exon splicing and the confidence with which we can be sure that a complete ASG has been recovered. We have also considered algorithms for calculating the minimum and maximum number of informative transcripts available from an ASG. Source code, as well as the statistical tests outlined and their results, are available from `http://www.stats.ox.ac.uk/~jenkins/ASG/`. Our method for testing the coverage of an ASG by its transcripts can provide experimentalists with a way to quantify any bias in the distribution of the transcriptome. In our examples we have been restricted to existing EST data, which can be somewhat limited both in quality and quantity. Quantitative analysis of the ASG will become far more fruitful when high-throughput microarray data on alternative splicing is more readily available, from which accurate probabilities can be associated with each splicing event. An important next step will then be to begin to incorporate knowledge of tissue-specific expression of particular isoforms, which has thus far been naïvely omitted.

Unfortunately most current microarray studies focus on individual splicing events—only 12.8% of alternative splicing relationships have been detected in full-length transcripts [10], but we envisage this to improve as the need to observe whole transcripts pushes the technology in this direction. When full-length transcripts are available, one way to look more closely at the conditional probabilities inherent in an ASG would be to focus on those transcripts revealing new edges to the ASG during sampling. The resulting 'signature' histogram can be compared to the same histogram generated by transcript simulation from one of the models, i.e. assuming no exon coupling (Fig. 4(*right*)). This figure and other of our tests suggest that a simple model for the distribution is Gaussian with mean between the minimum and maximum number of informative transcripts. Thus for example, strong positive correlation between exons would skew the distribution towards the minimum, compared to the distribution observed under the models. Across the 56 genes satisfying the pairwise model, the distribution of the mean number of informative transcripts reported was centred about 0.61 of the genes' ranges (i.e. between the minimum and maximum number of informative transcripts). All but 29 reported a mean in the range $(0.5, 0.7)$ and all but 7 were inside $(0.4, 0.8)$. For each gene the standard deviation in informative transcripts was less than 0.13 of the range.

6 Acknowledgements

Gil Ast and Richard Copley are thanked for their advice on which genes would be interesting. An anonymous reviewer is thanked for helpful comments. Thanks also to the LSI DTC at Oxford and to the EPSRC and BBSRC for its funding.

References

1. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., Shoemaker, D.D.: Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302** (2003) 2141–2144
2. Leipzig, J., Pevzner, P., Heber, S.: The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. Nucleic Acids Research **32** (2004) 3977–3983
3. Lareau, L.F., Green, R.E., Bhatnagar, R.S., Brenner, S.E.: The evolving roles of alternative splicing. Current Opinions in Structural Biology **14** (2004) 273–282
4. Black, D.L.: Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry **72** (2003) 291–336
5. Lopez, A.J.: Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annual Review of Genetics **32** (1998) 279–305
6. Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., Frey, B.J., Blencowe, B.J.: Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular Cell **16** (2004) 929–941
7. Heber, S., Alekseyev, M., Sze, S.H., Tang, H., Pevzner, P.A.: Splicing graphs and EST assembly problem. Bioinformatics **18** (2002) S181–188
8. Black, D.L.: A simple answer for a splicing conundrum. Proceedings of the National Academy of Sciences of the United States of America **102** (2005) 4927–4928

9. Ibrahim, E.C., Schaal, T.D., Hertel, K.J., Reed, R., Maniatis, T.: Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. Proceedings of the National Academy of Sciences of the United States of America **102** (2005) 5002–5007
10. Lee, C., Atanelov, L., Modrek, B., Xing, Y.: ASAP: the alternative splicing annotation project. Nucleic Acids Research **31** (2003) 101–105
11. Li, W.N., Reddy, S.M., Sahni, S.: On path selection in combinational logic circuits. IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems **8** (1989) 56–63
12. Lee, C., Roy, M.: Analysis of alternative splicing with microarrays: successes and challenges. Genome Biology **5** (2004) 231
13. Lee, C., Wang, Q.: Bioinformatics analysis of alternative splicing. Briefings in Bioinformatics **6** (2005) 23–33
14. Castle, J., Garrett-Engele, P., Armour, C.D., Duenwald, S.J., Loerch, P.M., Meyer, M.R., Schadt, E.E., Stoughton, R., Parrish, M.L., Shoemaker, D.D., Johnson, J.M.: Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biology **4** (2003) R66
15. Modrek, B., Lee, C.: A genomic view of alternative splicing. Nature Genetics **30** (2002) 13–19
16. Tabuchi, K., Südhof, T.C.: Structure and evolution of neurexins: insight into the mechanism of alternative splicing. Genomics **79** (2002) 849–859
17. Frank, N.Y., Margaryan, A., Huang, Y., Schatton, T., Waaga-Gasser, A.M., Gassser, M., Sayegh, M.H., Sadee, W., Frank, M.H.: ABCB5-mediated doxorubicin transport and chemoresistance in human malignant melanoma. Cancer Research **65** (2005) 4320–4333