**Drosophila GRAIL: An Intelligent System for Gene
Recognition in Drosophila DNA Sequences***

**Ying Xu, Gregg Helt†, J. Ralph Einstein,
Gerry Rubin†, and Edward C. Uberbacher**

**Informatics Group
Intelligent Systems Section
Engineering Physics and Mathematics Division
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, Tennessee 37831-6364**

**†Department of Molecular Biology
University of California at Berkeley
Berkeley, CA**

"The submitted manuscript has been
authored by a contractor of the U.S.
Government under contract No.
DE-AC05-84OR21400. Accordingly, the
U.S. Government retains a nonexclusive,
royalty-free license to publish or
reproduce the published form of this
contribution, or allow others to do so,
for U.S. Government purposes."

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

To be published as full paper with The First International IEEE Symposium on Intelligence in Neural and Biological Systems, May 23 - 24, 1995.

*Research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Drosophila GRAIL: An Intelligent System for Gene Recognition in Drosophila DNA Sequences

Ying Xu, Gregg Helt[†], J. Ralph Einstein,
Gerry Rubin[†] and Edward C. Uberbacher

Informatics Group
Computer Sciences and Mathematics Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364
and

[†]Department of Molecular and Cell Biology
University of California at Berkeley
Berkeley, CA 94704

Abstract

An AI-based system for gene recognition in Drosophila DNA sequences was designed and implemented. The system consists of two main modules, one for coding exon recognition and one for single gene model construction. The exon recognition module finds a coding exon by recognition of its splice junctions (or translation start) and coding potential. The core of this module is a set of neural networks which evaluate an exon candidate for the possibility of being a true coding exon using the "recognized" splice junction (or translation start) and coding signals. The recognition process consists of four steps: generation of an exon candidate pool, elimination of improbable candidates using heuristic rules, candidate evaluation by trained neural networks, and candidate cluster resolution and final exon prediction. The gene model construction module takes as input the clustered exon candidates and builds a "best" possible single gene model using an efficient dynamic programming algorithm. 129 Drosophila sequences consisting of 441 coding exons including 216358 coding bases were extracted from GenBank and used to build statistical matrices and to train the neural networks. On this training set the system recognized 97% of the coding messages and predicted only 5% false messages. Among the "correctly" predicted exons, 68% match the actual exon exactly and 96% have at least one edge predicted correctly. On an independent test set consisting of 30 Drosophila sequences, the system recognized 96% of the coding messages and predicted 7% false messages.

1. Introduction

Recognition of biologically relevant features in genomic sequences, such as coding regions, splice junctions and promoters, remains one of the major challenges of genome projects. With a number of sequencing projects planning sequencing rates of several megabases per year, the demands on effective computational technology for genomic feature recognition can

be expected to increase dramatically. Though the performance of feature recognition systems and methods is currently limited, this basic approach may soon be the only one capable of providing analysis at a rate compatible with the worldwide sequencing throughput.

This paper describes an AI-based system for locating coding exons (in the following, we simply call them *exons*) and building gene models in DNA sequences from *Drosophila melanogaster*. The system is developed using the general framework of the GRAIL gene recognition system [2, 5] developed at Oak Ridge National Laboratory, which was originally designed to recognize human genes. GRAIL finds an exon by recognizing its coding potential and its boundaries (splice junctions, translation start or stop codon). It achieves this by considering discrete exon candidates instead of a continuous coding probability function. An exon candidate is a translationally open region that is bounded by a putative acceptor junction (or translation start) and a putative donor junction (or a stop codon), and has an inherent reading frame. Each candidate is evaluated for coding potential using a window tailored to the region covered by the candidate, and its boundaries are evaluated for the possibility of being splice junctions or translation start. Using the "recognized" splice junction (or translation start) and coding signals, an exon candidate is scored for its degree of correctness as a true exon. By considering discrete candidates instead of using fixed-size sliding windows, as commonly used in coding region recognition systems, GRAIL's performance on exon recognition is almost exon size-independent [5].

GRAIL constructs a single "best" gene model in a specified DNA region using the scored exon candidates. A gene model can be considered as a series of non-overlapping exon candidates such that the adjacent exon candidates are reading-frame compatible with each other. GRAIL builds a gene model that con-

Figure 1: GRAIL gene recognition system

tains as many good-scoring exon candidates as possible from the pool of candidates.

Figure 1 illustrates the structure of the GRAIL gene recognition system. The system first generates all possible exon candidates that start with a putative translation start or acceptor junction and end with a putative donor junction or a stop codon. Four types of candidates are generated, i.e., starting, internal, terminal, and candidates that start with a putative translation start and end with a stop codon (we call them *single* exon candidates, corresponding to genes with only one exon). The system then eliminates the vast majority of the improbable candidates using a set of heuristic rules, each of which defines some necessary condition for a *true* exon. The surviving candidates are evaluated for the degree of correctness for being a true exon by a pre-trained neural network system. A clustering procedure divides the scored candidates into groups, each of which corresponds to one putative exon (but with different edge assumptions). The exon prediction (prior to gene modeling) is made by selecting the highest scoring candidate from each group. Using the clustered exon candidates, the GRAIL gene model construction module builds a "best" gene model by solving a constraint combinatorial optimization problem.

We have adopted this framework of GRAIL in designing a new system, *Drosophila* GRAIL, for *Drosophila* gene recognition. Under this framework, we have (1) implemented new algorithms to recognize translation starts, acceptor and donor splice junctions of *Drosophila* sequences, (2) revised the GRAIL coding recognition algorithms in accordance with the *Drosophila* sequence data, (3) developed a set of heuristic rules to define a *probable* exon candidate, and (4) designed and trained neural networks for exon recognition.

Coding potentials are measured by a hexamer frame-dependent preference algorithm [2, 3] and a non-

homogeneous 2nd order Markov chain-based algorithm [1, 3]. A total of 129 *Drosophila* sequences, consisting of 441 coding exons and 216358 coding bases, were extracted from GenBank (release # 82.0) and used to generate the statistical matrices for both algorithms. Splice junctions (and translation starts) are recognized using position-dependent dinucleotide statistical matrices. DNA segments from the flanking regions of splice junctions (or translation starts) were extracted from the 129 *Drosophila* sequences to generate these matrices.

Performance comparisons have been made between *Drosophila* GRAIL and the *Drosophila*-specific Genefinder program (Phil Green and L. Hillier, unpublished results), a gene prediction system that has been used extensively in the *C. elegans* genome project [4]. Genefinder uses several organism-specific tables in making its predictions. We constructed these tables for Genefinder using the same set of sequences used for training *Drosophila* GRAIL. On 30 independent test sequences extracted from GenBank (release # 84.0), *Drosophila* GRAIL compares favorably to Genefinder.

2. Methods

2.1. Exon Recognition Module

As in any pattern recognition problem, to recognize exons we need to find a set of features that are associated with exons, and to design an effective method to discriminate exons from non-exonic regions based on these features. Three types of information are used in our exon discrimination process. They are statistical measures of frequencies of different "vocabularies", like hexamers, in exons versus non-exonic regions; statistical measures of edge signals for splicing sites and translation starts; and measures of domain information including local G/C composition and the size of a candidate. The first type measures the coding potential of an exon candidate, the second measures the possibility of an exon candidate having correct boundaries, and the third provides information to the discrimination process in evaluating the significance of the coding scores and edge signal scores.

The exon recognition algorithm uses a hybrid artificial intelligence system which combines heuristic methods, neural networks and a clustering algorithm. The following gives the detail of the four main steps of the algorithm.

Step 1. Candidate pool generation

The algorithm takes as initial exon candidates any translationally open region greater than 11 bases long which begins and ends with minimal edge signals as defined below. It uses four simple enumeration procedures to generate a candidate pool of all possible starting, internal, terminal and single exon candidates. For starting exon candidates, a possible translation start and a possible donor junction are necessary for edge

signals (three bases *ATG* and two bases *GT* are the requirements for a possible translation start and a possible donor junction, respectively). For internal exon candidates, the minimal requirements are a possible splice acceptor (three bases *YAG* are the requirement) and a possible donor (*GT*). The minimal requirements for a terminal exon candidate are a possible acceptor junction (*YAG*) and a translation stop codon. A single exon candidate must have a possible translation start and a stop codon.

Several thousand candidates are typically generated for a sequence of 10 kb long.

Step 2. Feature recognition and filtering

The vast majority of the initial exon candidates are very unlikely to represent actual coding regions. We eliminate most of these using a set of simple heuristic rules, resulting in a much smaller and hence simpler set of candidates which the neural networks in Step 3 are trained to evaluate. Each of the rules is defined in terms of the "recognized" coding and edge signals, and defines some necessary conditions for a *probable* exon candidate. Failing any one of these conditions results in the elimination of the candidate.

Step 2.1. Feature recognition

Two types of features are recognized in the system. They are the coding potential of a candidate region and the edge signals of a candidate.

Coding potential recognition

The coding potential of a candidate is measured by two methods. They are the frame-dependent hexamer preference method [2, 3] and a 2^{nd} order non-homogeneous Markov chain model [1, 3]. The former measures the coding potential by comparing the *a priori* probabilities of the hexamers appearing in coding regions and non-coding regions while the latter, in essence, measures the coding potential similarly but using trimers and based on different hypotheses about the DNA sequences (Markov model treats a DNA sequence as a stochastic process). For each candidate, the two methods measure its coding potential using a single window tailored to the candidate region.

To fully utilize the information available under this scheme, we also measure the coding potential of the two 50-base regions adjacent to a candidate. The rationale is that a true exon should receive high coding scores within it but low scores in the regions surrounding it. This extra piece of information not only helps to eliminate false candidates, but also helps to measure the degree of correctness of a partially correct candidate in Step 3.

Splice junction recognition

Splice acceptor and donor junctions are recognized by a simple position-dependent dimer preference algorithm, respectively. The algorithm evaluates every possible splice junction (satisfying the minimum requirements for a splice junction given in Step 1) for the possibility of being a true splicing site, using a dimer preference matrix obtained from the flanking regions of all possible splicing sites. For acceptor junctions, segments of 25-bases long surrounding all *YAG*'s from position -20 to position +4 (the *YAG* is in positions 0, +1, +2) are extracted. For donor junctions, segments of 25-bases long surrounding all *GT*'s from position -4 to position +20 (the *GT* is in position 0, +1) are extracted. The normalized frequencies of each of the 16 possible dimers from the true and false splice junction samples, for acceptors and donors, are used to calculate the preference matrices.

Potential splice acceptor junctions $a_{-20}...a_{+4}$ are evaluated using the following function

$$P_{acc}(a_{-20}...a_{+4}) = \sum_{i=-20}^0 P_i(a_i a_{i+1}) + \sum_{i=3}^4 P_i(a_i a_{i+1}),$$

and similarly potential donor junctions $a_{-4}...a_{+20}$ are evaluated using the following

$$P_{don}(a_{-4}...a_{+20}) = \sum_{i=-4}^{-1} P_i(a_i a_{i+1}) + \sum_{i=2}^{20} P_i(a_i a_{i+1}),$$

where $P_i(a_i a_{i+1})$ denotes the *preference value* of dimer $a_i a_{i+1}$ (the log ratio of frequencies of $a_i a_{i+1}$ appearing in the i^{th} position of a splice junction region versus in bulk DNA) of in position i .

Translation start recognition

The translation start recognition algorithm is a 2-step process. It evaluates the potential translation starts (an *ATG* is the minimum requirement) for the possibility of being a true translation start by first filtering out the *improbable* candidates and then evaluating the remaining ones with a neural network. A position-dependent dimer preference algorithm similar to the above is used to do the filtering. Segments of 15-bases long surrounding all *ATG*'s from position -7 to position +7 (the *ATG* is in positions 0, +1, +2) are extracted. The normalized frequencies of each of the 16 possible dimers from the true and false translation start samples are used to calculate the preference matrix.

For each potential translation start site $a_{-7}...a_{+7}$, the following function is used to measure its possibility of being a true translation start:

$$P(a_{-7}...a_{+7}) = \sum_{i=-7}^{-1} P_i(a_i a_{i+1}) + \sum_{i=3}^7 P_i(a_i a_{i+1}),$$

where $P_i(a_i a_{i+1})$ is the preference value of dimer $a_i a_{i+1}$ in position i . Threshold $P() \geq 0$ is used to filter out "false" translation starts.

The remaining translation start candidates are evaluated by a trained neural network. The neural net takes the $P_i(a_i a_{i+1})$'s as input and outputs a value between 0 and 1, indicating the possibility as a true translation start. The net has 12 input nodes, a hidden layer with 2 nodes, and one output node. The neural network was trained using 136 true and over 1000 false translation starts.

Step 2.2. Application of heuristic rules

A total of 15 rules are currently used in the system. All the rules are derived from existing knowledge and the statistical analysis of the training set.

On average, application of the heuristic rules eliminates over 95% of the initial candidates. About 2 - 3% of the true exons are lost at this step. Among the candidates that passed the rules, only about 15% are false candidates. As a result of applying the rules, the exon discrimination problem for the neural networks in Step 3 is greatly simplified.

Step 3. Candidate evaluation

Recognized features are combined by a set of neural networks in evaluating a candidate for possibility of being a true exon. The exon evaluation algorithm uses four feed-forward neural networks, which have been trained using a back-propagation algorithm, to evaluate the *degree* of correctness of an exon candidate, for each type of exon, respectively. In the current implementation, we use the following features associated with a candidate in the neural network evaluation (we use the internal exon candidate evaluation as an example; solutions to the other cases are similar).

- **Coding signals:** A total of six coding scores are used in the neural network evaluation. They are frame-dependent hexamer scores for the candidate and the two 50-base regions surrounding it, and the non-homogeneous 2^{nd} order Markov chain scores for the candidate and its two 50-base surrounding regions.
- **Edge signals:** Both the potential acceptor and donor scores obtained above are used in the neural network evaluation process.
- **Supporting information:** Some extra information is used in the evaluation process mainly to help in evaluating the significance of above scores. These bias-correction parameters include the G/C composition of a 2,000-base region centered around the candidate, the G/C composition of the candidate itself, and the size of the candidate.

As stated in Step 2, most of the candidates that have passed the heuristic rules are partially correct, i.e., they overlap the corresponding exons in the same translation frames. So the main goal of the neural network evaluation is to score the *partial correctness*

(defined by the following matching function) of a candidate. To train the neural network, a simple matching function $M()$ is used to represent the correspondence of a given candidate with the actual exon(s):

$$M(\text{candidate}) = \left(\sum_i m_i / \text{size}(\text{candidate}) \right) \times \left(\sum_i m_i / \sum_j \text{size}(\text{exon}_j) \right),$$

where $\sum_i m_i$ is the total number of bases of the candidate that overlap some actual exons (in the same translation frame), and $\sum_j \text{size}(\text{exon}_j)$ is the total length of all the exons that overlap the candidate. Using such a function helps "teach" the neural network to discriminate between candidates with different degrees of overlap with actual exon(s).

The neural network for evaluating internal exon candidates is a feed-forward network with two hidden layers. It has 11 input nodes and 1 output node. The first hidden layer has 7 nodes and the second one has 3. The network is trained to score a given set of 11 parameters between 0 and 1, with 1 intending to represent a perfect match with a true exon. The system is trained using a total of about 2,000 actual exons, partially correct exon candidates and false candidates.

Step 4. Clustering of candidates and exon prediction

The candidates scored by the neural networks form a set of "clusters" of overlapping candidates. In the ideal situation, each cluster would correspond to one actual exon. However, as shown in Figure 2(a), in sequences with long open reading frames, one cluster may cover a region corresponding to more than one exon. The goal of this step is to divide such a "cluster" into smaller groups so that each group has a better one-to-one correspondence with a single exon.

Figure 2(b) shows that centers of the candidates form groups that have better one-to-one correspondence with actual exons. We have implemented the following algorithm that separates each natural "cluster" into one or more groups using the centers of the candidates. The algorithm has the following two steps.

- **Separate Step:** Divides each "natural" cluster into smaller groups so an objective function is optimized.
- **Merge Step:** Selects the highest scoring candidate from each group, and merges any groups which have their highest scoring candidates overlapping.

The Separate Step divides candidates' centers into groups so that the distance between two adjacent groups is "significantly" larger than the average distance between adjacent centers within each group, and the total number of groups formed from the "cluster" is "reasonably small". Specifically, it uses two application-specific parameters R and G , and guarantees the ratio of the distance between two adjacent groups and the average distance between two adjacent

Figure 2: Neural network output. The solid bars on top represent real exons. The x-axis represents the DNA sequence from left to right. The rectangles in the middle are the predicted exon candidates, where the height of a rectangle indicates the score of the candidate. The vertical lines in the bottom are the centers of the corresponding rectangles. Sequence: DMSUWA

centers within each group to be bigger than R , and the number of partitioned groups to be less than G . The algorithm finds a partition of the "cluster" that satisfies these conditions and furthermore minimizes the sum of the average distances between two adjacent centers of all groups.

After clustering, the system selects the highest scoring candidate from each cluster as its prediction of the putative exon (prior to gene modeling).

2.2. Gene Model Construction Module

The gene model construction module builds a single gene model in a specified DNA region using the predicted exon candidates. The goal of the gene model construction is to linearly append the predicted exons in such a way that a series of constraints are satisfied. These constraints include the following: (1) adjacent exons in the gene model are reading-frame compatible to each other; (2) the distance between two adjacent exons is bigger than some given constant – the minimum intron size; (3) no in-frame stop codons are formed when appending two adjacent exons; (4) the gene model optimizes a pre-defined objective function.

Statement of the problem

We first give a few definitions. For an exon E , $\text{left}(E)$ and $\text{right}(E)$ denote the indices of E 's left and right edges, respectively. $\text{frame}(E)$ denotes the reading frame of E . Each exon E has a non-negative score $p(E)$, which is the neural net score of E . Two non-overlapping exon candidates E_1 and E_2 , with E_1 preceding E_2 , are said to be *reading-frame compatible* with each other if there is a putative donor to the right of E_1 and a putative acceptor to the left of E_2 , and

$$\text{frame}(E_2) = (\text{left}(E_2) - \text{right}(E_1) + \text{frame}(E_1)) \bmod 3.$$

A series of non-overlapping exons $E_1, E_2, \dots, E_n$, with E_i preceding E_{i+1} , forms a *partial gene model* if (1)

E_i and E_{i+1} are reading-frame compatible, (2) no in-frame stop codon is formed when appending E_i and E_{i+1} , (3) $\text{left}(E_{i+1}) - \text{right}(E_i) \geq K$, where K is the minimum intron size and has value 50 in *Drosophila* GRAIL.

Let $\{C_1, C_2, \dots, C_m\}$ be a set of clusters and C_i contain a number of exon candidates. Our goal is to select a set of non-overlapping exon candidates $E_1, E_2, \dots, E_n$, at most one from each cluster, to form a partial gene model that maximizes the following function:

$$\sum_{i=1}^n p(E_i) - \sum_{i=1}^m \mathcal{P}(C_i) - \mathcal{P}_s(E_1) - \mathcal{P}_t(E_n),$$

where $\mathcal{P}(C_i)$ is a penalty factor, which has maximum score of the exons from cluster C_i if no exon from C_i is included in the gene model otherwise it is zero; $\mathcal{P}_s$ and $\mathcal{P}_t$ are fixed positive values when the gene model is missing the starting exon and terminal exon, respectively, otherwise they are zeros.

Dynamic programming algorithm

A dynamic programming algorithm is used to solve the optimization problem defined above. The algorithm scans through clusters from left to right until it exhausts all the clusters.

For each cluster, the algorithm builds (at most) 18 best (in the sense of maximizing the above objective function) partial gene models that end with exon candidates from this cluster, based on the best partial gene models which end with exon candidates of the previous clusters, counted from left to right. When extending a current partial gene model to the right to include one more exon, the algorithm checks if the conditions for a partial gene model are satisfied. By doing so, the algorithm builds the best partial gene model in a time proportional to the product of the total number of exons and the number of clusters.

To check if the conditions for a partial gene model are satisfied when extending gene models from left to right, some information needs to be provided about the reading frames and ending edges of (the last exons of) the previous models. We do this as follows. For each cluster under consideration, we construct a best partial gene model that ends with an exon candidate of this cluster for each of the following situations. The exon candidate could be in reading frame $\alpha \in \{0, 1, 2\}$ and its right edge modulo 3 could be in any codon position $\beta \in \{0, 1, 2\}$. To also take into consideration the possibility of forming an in-frame stop codon when appending two adjacent exons, we distinguish the following situations for each possible α and β . Let

$$\mathcal{D} = (\alpha + 3 - \beta) \bmod 3.$$

When $\mathcal{D} = 0$ the exon could end with a *T* or non-*T* letter, and when $\mathcal{D} = 1$, the exon could end with bases *TA*, *TG*, or any other doublet. It is shown [6] that considering all these 18 possible situations provides sufficient and necessary information for the gene model construction algorithm. We refer the reader to [6] for more details of the dynamic programming algorithm.

Figure 3 lists four examples of the *Drosophila* GRAIL exon recognition and gene modeling modules.

3. Results and Discussion

129 *Drosophila* sequences were extracted from the GenBank (release # 82.0) to build the statistical matrices and to train the neural networks. On the training set the system has located 90% of the exons and 97% of the coding message, with 7% falsely predicted exons and 5% falsely predicted message. On an independent test set (GenBank release # 84.0) containing 30 *Drosophila* sequences, the performance of the system is statistically similar to the one on the training set. It recognizes 90% of the exons and 96% of coding message with 6% falsely predicted exons and 7% false message.

The performance of the system is evaluated in greater detail below. Table I summarizes the performance of the splice junction recognition algorithms on the 129 training sequences. Each possible splice junction is scored and the score is normalized into a range from -10 to +10.

where # and % denote the total number in that category and percentage of the number with respect to the total number, respectively.

As can be seen from Table I, by using a threshold 0, the splice junction recognition algorithms eliminate 96% (23265/24160) of the false donors and 90% (11358/12253) of the false acceptors from the initial splice junction candidate pool. Among the 1188 donor candidates that pass the threshold 293 (25%) are real donors, and among the 1889 acceptor candidates 284 (15%) are real. Though these numbers show that we are still far from solving the splice junction recognition problem much better recognition is achieved during the definition of probable exon candidates, which

Table I

	Donors				Acceptors			
	True		False		True		False	
	#	%	#	%	#	%	#	%
+10	87	28	26	0	104	35	110	0
+9	32	10	5	0	34	11	50	0
+8	32	10	16	0	33	11	84	0
+7	32	10	55	0	27	9	103	0
+6	34	11	47	0	22	7	153	1
+5	31	10	87	0	28	9	205	1
+4	14	4	138	0	10	3	229	1
+3	17	5	207	0	6	2	282	2
+2	11	3	314	1	12	4	389	3
+1	3	0	374	1	6	2	402	3
0	5	1	1147	4	4	1	999	8
-1	1	0	820	3	0	0	571	4
-2	2	0	1067	4	1	0	614	5
-3	0	0	1262	5	1	0	683	5
-4	1	0	1362	5	0	0	678	5
-5	2	0	1591	6	1	0	706	5
-6	0	0	1697	7	0	0	696	5
-7	0	0	1844	7	0	0	650	5
-8	1	0	1823	7	1	0	637	5
-9	0	0	1850	7	0	0	591	4
-10	0	0	8423	34	1	0	3421	27

must contain two good splice junctions (or a translation start or stop) and have a strong coding potential. Among the located exons 68% have both their edges correct and 96% have at least one edge correct.

Table II gives the prediction results of the first step of this 2-step process. The neural network in the second step improves upon this slightly.

Table III summarizes the performance of the exon recognition algorithm on the 129 training sequences. Exons are divided into four categories: starting, internal, terminal and single exons. Among the 115 starting exons, 21 of them are 20 bases or shorter, which may have also contributed to the relatively poor performance of the system on the starting exons.

The gene model construction module builds a single gene model from a given set of cluster of exon candidates. In cases where a DNA sequence has more than one gene we have manually divided the sequence into pieces so that each piece contains one gene. Table IV summarizes the performance of the gene model construction module on the combined set of test and training sequences. As can be seen from the table, the false positive rates went down and the edge accuracy went up significantly while the true negative rate went up compared to the prediction of the exon recognition module. This is the result of mainly two things. The enforcement of reading-frame compatibility in a gene model reduces the false positive rate but also slightly reduces the true positive rate. The main cause for the

Figure 3: Prediction and gene modeling results on sequences (A) DROLAMB2A, (B) DROSEV, (C) DROOTUA and (D) DROSUSG. For each of the four predictions, the solid bars on top are the real exons. The solid bars connected by lines in the middle are the predicted gene model. The rectangles in the bottom represent the exon predictions, with heights proportional to their scores.

Table II

	True starts		False starts	
	#	%	#	%
+10	60	41	347	4
+9	9	6	131	1
+8	7	4	143	1
+7	10	6	143	1
+6	10	6	180	2
+5	11	7	216	2
+4	4	2	279	3
+3	2	1	242	3
+2	7	4	310	3
+1	5	3	311	3
0	6	4	737	9
-1	2	1	366	4
-2	5	3	431	5
-3	1	0	388	4
-4	2	1	418	5
-5	0	0	356	4
-6	0	0	407	5
-7	0	0	355	4
-8	0	0	326	4
-9	2	1	301	3
-10	3	2	1473	18

reduction of true positive rate in the gene model prediction is from the incorrect exon type prediction, e.g., a starting exon may be predicted as an internal exon due both to a weak translation start signal and a false acceptor signal in the 5' region of the exon. This type of false prediction results in the two predicted exon candidates being not reading-frame compatible.

Comparisons have been conducted between Drosophila GRAIL exon recognition module and Genefinder (the Drosophila-specific version). Table V summarizes the performance of the two programs on the 129 training sequences and 30 test sequences, respectively. The training set contains 441 exons, 216358 coding bases and 271178 non-coding bases, and the test set contains 104 exons, 47083 coding bases and 64391 non-coding bases. We have set some of the Genefinder's parameters in such a way that the true positive rate of its prediction is comparable to Drosophila GRAIL's prediction. As can be seen from Table V, Genefinder gives a higher false positive rate when the two systems true positive rates are about the same.

In conclusion, we have used the framework of the GRAIL gene recognition system to build a new system for recognizing genes on Drosophila sequences. The preliminary test results indicated the potential useful-

Table III

	# Exons	# Exons found	%	# Bases	# Bases found	%
Starting	115	92	80%	33401	31357	94%
Internal	190	184	97%	93780	92295	98%
Terminal	115	100	87%	64964	61491	95%
Single	21	19	90%	24213	23754	98%
Total	441	395	90%	216358	208897	97%

Table IV

Prediction Modeling	# Exons	# F. exons	# Bases	# F. bases	# 1-edge	# 2-edge
	489 (90%)	39 (7.4%)	254098 (96%)	13257 (5%)	468 (96%)	331 (68%)
	478 (88%)	16 (3%)	250405 (94%)	5927 (2%)	464 (97%)	382 (80%)

Exons and # Bases are the correctly found exons and coding bases, respectively, # F. exons and # F. bases are the falsely-predicted exons and coding bases, and # 1-edge and # 2-edge are the correctly predicted exons with one edge and both edges correct, respectively.

Table V

	# Exons found	# False exons	# Bases found	# False bases	CC ¹
Genefinder	392 (89%)	81 (17%)	207980 (96%)	20546 (9%)	0.882
	94 (90%)	20 (18%)	45200 (96%)	5903 (12%)	0.861
DGRAIL	395 (90%)	33 (7.7%)	208897 (97%)	9910 (5%)	0.928
	94 (90%)	6 (6%)	45201 (96%)	3347 (7%)	0.905

¹ CC is the correlational coefficient, and is calculated using the following formula

$$CC = \frac{B_n \times TP - B_c \times FP}{\sqrt{B_c \times B_n \times (TP + FP) \times (B_c + B_n - TP - FP)}}$$

where B_c and B_n are the numbers of coding and non-coding bases, respectively, and TP and FP are number correctly predicted and falsely predicted coding bases, respectively.

ness in locating genes in anonymous *Drosophila* DNA sequences.

The system will be available through both an e-mail server and a client/server program XGRAIL at ORNL. For more information send e-mail to grail@mail@ornl.gov.

Acknowledgments

This research was supported by the Office of Health and Environmental Research, United States Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

References

- [1] M. Borodovsky, Yu. Sprizhitskii, E. Golovanov and A. Aleksandov, "Statistical Patterns in the Primary Structures of Functional Regions in *E. Coli*," *Molekulyainaya Biologiya*, 20, pp. 1390 - 1398, 1986.
- [2] E. C. Uberbacher and R. J. Mural, "Locating protein-coding regions in human DNA sequences by a multiple sensors-neural network approach", *Proc. Natl. Acad. Sci. USA*, Vol. 88, pp. 11261 - 11265, 1991.
- [3] E. C. Uberbacher, J. R. Einstein, X. Guan, and R. J. Mural, "Gene recognition and assembly in the GRAIL system: progress and challenges", *Proceedings of The 2nd International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, H. A. Lim, J. W. Fickett, C. R. Cantor and R. J. Robbins, Eds, World Scientific, pp. 465 - 476, 1993.
- [4] R. Wilson, *et al.*, "2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*", *Nature*, Vol. 368, pp. 32 - 38, 1994.
- [5] Y. Xu, R. J. Mural, M. Shah, and E. C. Uberbacher, "Recognizing exons in genomic sequence using GRAIL II", *Genetic Engineering: Principles and Methods*, Jane Setlow, Ed., Plenum Press, Vol. 16, pp. 241- 253, July 1994.
- [6] Y. Xu, R. J. Mural and E. C. Uberbacher, "Constructing gene models from accurately-predicted exons: an application of dynamic programming", *Computer Applications in the Biosciences*, Vol. 10, No. 6, pp. 613 - 623, 1994.

CONF-9505100--1

Note: This is a draft of a paper submitted for publication. Contents of this paper should not be quoted or referred to without permission of the author(s).

To be presented at the 4th European Microbeam Analysis Society Workshop
on Modern Developments and Applications in Microbeam Analysis,
St. Malo, France, May 14-19, 1995
Published in *Proceedings*

MICROANALYSIS AT ATOMIC RESOLUTION

S. J. Pennycook, D. E. Jesson, N. D. Browning,* and M. F. Chisholm
Solid State Division, Oak Ridge National Laboratory
P.O. Box 2008, Oak Ridge, Tennessee 37831-6030

*Department of Physics (M/C 273), University of Illinois at Chicago
845 West Taylor Street, Chicago, Illinois 60607-7059

"The submitted manuscript has been authored
by a contractor of the U.S. Government under
contract No. DE-AC05-84OR21400.
Accordingly, the U.S. Government retains a
nonexclusive, royalty-free license to publish or
reproduce the published form of this
contribution, or allow others to do so, for U.S.
Government purposes."

Prepared by
SOLID STATE DIVISION
OAK RIDGE NATIONAL LABORATORY
Managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
under
Contract No. DE-AC05-84OR21400
with the
U.S. DEPARTMENT OF ENERGY
Oak Ridge, Tennessee

April 1995

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

JK

MASTER

Microanalysis at Atomic Resolution

S. J. Pennycook, D. E. Jesson, N. D. Browning* and M. F. Chisholm

Solid State Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6030

Abstract

In this paper we discuss how the dedicated scanning transmission electron microscope can provide a microanalysis of crystalline materials at atomic resolution. The method requires the establishment of incoherent conditions for a reference imaging signal as well as the spectroscopic signal. The image can then be used to focus and locate the probe to atomic precision for microanalysis. The Z-contrast image provides the most convenient incoherent reference image, and X-ray and electron energy loss data may be acquired simultaneously. In zone axis crystals, strong columnar channelling delays the onset of beam broadening for several hundred Ångströms, so that atomic resolution microanalysis may be achieved in materials specimens of significant thickness. This combination of signals provides a powerful means for studying interface structure and bonding, and avoids relying on preconceived model structures.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

* Present address: Department of Physics (M/C 273), University of Illinois at Chicago, 845 West Taylor Street, Chicago, IL 60607-7059. USA

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

1. Introduction

Now that the dedicated scanning transmission electron microscope (STEM) allows atomic sized probes to be routinely achieved, the potential to analyse chosen atomic columns for their composition and bonding has become a reality. Studies of materials can finally be carried out at the fundamental atomic scale. For example, impurity segregation at dislocation cores in grain boundaries can be investigated atomic column by atomic column. Such capabilities are expected to have a far-reaching impact throughout materials science [1].

In this paper we discuss the conditions under which such microanalysis is possible, including considerations of probe formation, beam broadening, and localization of spectroscopic information. The first critical requirement is the ability to form an incoherent image of the material at atomic resolution, so that the probe size and position may be accurately controlled. The Z-contrast image formed from elastically scattered electrons with a high angle annular detector is convenient for this purpose, and does not interfere with the simultaneous acquisition of X-ray or energy loss data. Secondly, incoherent conditions for the analytical signal must be established, so that it too will show an incoherent image in direct correspondence to the Z-contrast image, although typically of much weaker intensity. In this way, the Z-contrast image can provide a high intensity reference image corresponding to the much weaker inelastic image, and can therefore be used to focus the probe to an optimum size and to locate it over particular columns of interest. A point microanalysis can therefore be achieved from specific atomic columns, in principle from a single column, provided beam damage is not too rapid.

It is important to appreciate from the outset that atomic resolution microanalysis is only possible at present on the dedicated STEM, although there appears no reason in principle why this should not be achievable in a conventional TEM equipped with a field emission gun. The instrumental requirements are discussed further in the following section, while in section 3. we discuss the propagation of a coherent probe through a zone axis crystal. It is the remarkable behaviour of such a probe that makes atomic resolution microanalysis possible in specimens up to several hundred Ångstroms in thickness. In section 4. we discuss the limits on spatial resolution imposed by the delocalization of the inelastic scattering probability. Generally, losses in excess of several hundred eV are required for atomic resolution to be possible, although some improvement is possible by using an annular collector aperture to exclude low-angle scattering from the spectrometer.

Finally, in section 5 we present an experimental demonstration of atomic resolution microanalysis by mapping the Co L loss intensity across an atomically abrupt CoSi₂/Si(111) epitaxial interface. This intensity drops dramatically on moving from the last plane of the silicide to the first plane of the silicon, a distance of only 2.7Å.

2. Instrumental Requirements for Incoherent Imaging

As realized first by Crewe [3,4], the high brightness of the field emission gun is essential in achieving an atomic sized probe with sufficient current for image formation. To achieve an electron-optically limited probe, a high source demagnification is used, so that the geometrical size of the source in the focal

plane of the objective lens is much less than the actual probe size obtained by balancing the effects of spherical aberration and defocus. Under these conditions the probe is coherent, with an optimum size given by

$$d_{\text{opt}} = 0.43 C_s^{1/4} \lambda^{3/4} \quad (1)$$

where C_s is the objective lens spherical aberration coefficient, and λ the electron wavelength. This optimum probe is obtained with an optimum objective aperture

$$\alpha_{\text{opt}} = \left(\frac{4\lambda}{C_s} \right)^{1/4} \quad (2)$$

and an optimum defocus of

$$\Delta f_{\text{opt}} = -(C_s \lambda)^{1/2}. \quad (3)$$

These conditions were first identified by Scherzer [5], and later analysed in more detail by Crewe and coworkers [6,7]. Note that they differ from the more familiar Scherzer conditions for bright field phase contrast imaging. In particular, the optimum objective aperture tends to be smaller, since there is no need to include particular low order diffracted beams, and d_{opt} , which is often taken as a convenient measure of microscope resolution, is significantly improved over the corresponding expression for coherent imaging,

$$d_{\text{coh}} = 0.66 C_s^{1/4} \lambda^{3/4}. \quad (4)$$

The higher resolution is a direct result of squaring the probe amplitude profile $P(R)$, on which the resolution of a *coherent* image depends, to form

the intensity profile $P^2(R)$, which controls the resolution of an incoherent image or analysis. Probe profiles are straightforward to calculate [8], and Fig. 1. shows profiles as a function of defocus for the two STEM instruments at Oak Ridge National Laboratory, a 100 kV VG Microscopes HB501UX, with an objective lens C_s of 1.3 mm giving an optimum probe size of $2.2\text{\AA}$, and a 300 kV VG Microscopes HB603, with a 1 mm C_s giving a $1.3\text{\AA}$ probe. These profiles show behaviour typical of coherent probes, broad at low defocus values, becoming narrower at higher defocus but also acquiring significant subsidiary maxima. The Scherzer optimum probe provides the narrowest central peak without substantial intensity appearing in subsidiary maxima. This represents the most local probe for imaging and microanalysis.

The conditions required for incoherent imaging with elastically (including thermally) scattered electrons, and inelastically scattered electrons, have been reviewed recently [9] and will only be briefly mentioned here. Incoherent conditions apply when either *all* the scattering is detected corresponding to some elastic or inelastic interaction, or when a representative *fraction* of the total scattering is detected.

In the case of elastically scattered electrons incoherent imaging is achieved by collecting a small representative fraction of the total scattering through the use of a high angle annular detector. By ensuring that the inner angle θ_i of the detector is large compared to the probe forming aperture, incoherent imaging of crystal columns is achieved down to spacings of

$$\Delta R = 1.22 \lambda / \theta_i . \quad (5)$$

This holds even if the scattering reaching the detector is predominantly coherent [11], since the detector averages over the diffraction pattern. In most cases however, thermal diffuse scattering is the dominant means of scattering to the annular detector. Thermal vibrations destroy the long-range positional correlations within each column, so that each atom sees only a few neighbours within its column as correlated [12]. The length of such a "packet" depends on detector angle and Debye-Waller factor, but is generally less than the thickness of a typical specimen. In this case the intensity will be dependent simply on the number of packets in the column, ie. proportional to thickness t (ignoring for the moment effects due to channelling and absorption which are discussed in the next section). Columns that are much shorter than this correlation length will scatter predominantly coherently, so that the intensity will be proportional to t^2 [11].

With X-ray microanalysis, clearly we again collect a small but representative fraction of the total emission, but for electron energy loss spectroscopy (EELS), the strongly forward-peaked scattering makes it experimentally feasible to collect most of the scattering with modern aberration corrected spectrometers. Incoherent imaging conditions are again fulfilled [10], and the inelastic image is given by a convolution of the probe intensity profile with an inelastic object function centered on the atomic sites. However, because of the low characteristic scattering angles associated with low energy edges, the corresponding object functions may be substantially broader than the probe profile, so that to achieve atomic resolution a certain minimum energy loss is necessary as discussed in section 4.

It is clear that the conditions for incoherent imaging with elastically and inelastically scattered electrons are very complementary, so that both may be achieved simultaneously with the STEM geometry. The intensity in the electron optically limited probe is however significantly reduced from the conditions generally used for microanalysis, so that use of an efficient parallel detection system for EELS is essential. The Oak Ridge system uses a McMullan detector [13], based on a Wright Instruments multi-phase-pinned CCD camera, which provides exceptional sensitivity. Indeed, calculations indicate that single atom detection is feasible with this system [14], despite the low beam currents, due to the fact that we are illuminating only a few tens of atoms in one atomic column. A single atom therefore represents a much higher concentration than in situations where beam broadening is important, such as glasses or biological specimens [15]. Due to the likelihood of beam damage affecting the spectral fine structure, an important advantage of the McMullan system is its ability to record a sequence of spectra in rapid succession. Comparing the first spectrum with the last enables an assessment to be made of the importance of beam damage, and if there is none, the spectra may be summed to improve statistics. At present, only the 100 kV STEM is equipped with this EELS capability.

Although no X-ray microanalysis has yet been demonstrated at atomic resolution, the only limitation would appear to be the need for longer counting times with the increased possibility of beam damage. Although the EELS fine structure is the most sensitive indicator of beam damage, amorphization of a column is quite apparent from the Z-contrast image alone, and it would be reasonable to assume that if the image were not changed at the end of the analysis, then at least the major constituents of the

column were not displaced during the analysis. X-ray analysis requires the use of the virtual objective aperture, and due to the high demagnification of the illumination system, smaller apertures than usual are required. The optimum angle can be adjusted by varying the condenser and objective lens strengths; extra demagnification can be achieved if necessary by increasing the strength of the gun lens. In this way, Z-contrast images can be obtained of equal quality to those obtained with the real objective aperture.

3. Probe Channelling

No consideration has so far been given to the effects of dynamical diffraction on the conditions needed to establish incoherent imaging. The reason for this is that dynamical diffraction manifests itself in a particularly simple way when a coherent probe is incident on a zone axis crystal. It has the effect of focussing the current onto the atomic columns, while maintaining the overall envelope of the incident probe, as indicated schematically in Fig. 2. For an incident probe size below the column spacing, most of the current can be channelled along a single atomic column, which is critical for Z-contrast imaging and microanalysis at atomic resolution. Incoherent imaging is no longer restricted to weak phase objects, but applies also to crystals of substantial thickness.

This behaviour was apparent in early calculations of STEM probe propagation [16,17], and its explanation lies in the effects of quantum mechanical superposition [18-20]. The STEM probe is a coherent superposition of plane waves emanating from the objective aperture, and is best thought of as a spherical wave converging on the specimen surface. Inside the crystal, the

$$O(\mathbf{R}) = \sum_i \sigma_i(\mathbf{R}) \epsilon^{1s^2} b^{1s^2}(\mathbf{R}) \left(\frac{1 - e^{-2\mu^{1s}t}}{2\mu^{1s}t} \right) \quad (6)$$

where the sum is over all atoms i in the column, b^{1s} is the 1s Bloch state of excitation ϵ^{1s} and absorption coefficient μ^{1s} , calculated for exact zone axis incidence. The angular dependence of these quantities is included by defining an effective probe intensity profile,

$$P_{\text{eff}}^2(\mathbf{R}) = \left| \frac{1}{\epsilon^{1s}(0)} \int_{\text{objective aperture}} \epsilon^{1s}(\mathbf{K}) e^{i[\mathbf{K} \cdot (\mathbf{R} - \mathbf{R}_0) + \gamma(\mathbf{K})]} d\mathbf{K} \right|^2 \quad (7)$$

where $\mathbf{K}$ is the transverse component of the incident wave vector and

$$\gamma = \frac{\pi}{\lambda} \left(\Delta f \theta^2 + \frac{1}{2} C_s \theta^4 \right) \quad (8)$$

is the transfer function phase factor for a defocus Δf and electron wavelength λ . If the Bloch wave excitation ϵ^{1s} was constant over the aperture, this would be identical to the incident probe profile. Typically, ϵ^{1s} falls to about half its axial value at the edge of the objective aperture, which has the effect of broadening the effective probe by about 10% [19]. It can be appreciated that the constancy of ϵ^{1s} over the objective aperture is critical to forming a probe inside the crystal. ϵ^{2s} for example changes sign across the aperture, which is the reason that the 2s Bloch state, although highly excited on axis, does not contribute significantly to the probe intensity inside the crystal. The Z-contrast, EELS or X-ray image is then given by the convolution of the effective probe with the columnar cross section, which can be considered as the object function for incoherent imaging:

$$I(R_0) = \int P_{\text{eff}}^2(R-R_0) O(R) dR. \quad (9)$$

4. Localization of Inelastic Scattering

Object functions for elastic scattering to the high angle detector are highly localized about the atomic columns, but this is not necessarily the case for inelastic scattering. Plasmons and low loss excitations may be quite delocalized, and would not carry atomic resolution chemical information. (Note that atomic resolution images in plasmon losses can be obtained through elastic scattering mechanisms, a situation usually referred to as preservation of contrast. This is unrelated to the present discussion which is concerned with the localization of the microanalytical information.) A good description of the object function for inelastic processes is therefore desirable. The object function for a transition from an initial state $|0\rangle$ to a final state $|n\rangle$ may be written in terms of the matrix elements ρ_{no} as

$$O(R) = \left(\frac{e^2}{\pi \hbar v} \right)^2 \left| \int \frac{\rho_{no}(\mathbf{q})}{q^2} e^{-i\mathbf{K} \cdot \mathbf{R}} d\mathbf{K} \right|^2, \quad (10)$$

[10,21] where the momentum transfer $\mathbf{q} = (q_{z\min}, \mathbf{K})$, with $q_{z\min}$ corresponding to the minimum momentum transfer at zero scattering angle. The object function can also be written in terms of a form factor in reciprocal space

$$O(R) = \frac{1}{(2\pi)^2} \int f'(\mathbf{K}) e^{i\mathbf{K} \cdot \mathbf{R}} d\mathbf{K}. \quad (11)$$

where

$$f'(\mathbf{K}) = \left(\frac{e^2}{2\pi^2 \hbar v} \right) \int \frac{\rho_{no}(\mathbf{q}')}{q'^2} \frac{\rho_{no}^*(\mathbf{q}' + \mathbf{K})}{|\mathbf{q}' + \mathbf{K}|^2} d\mathbf{K}' \quad (12)$$

Although this form factor has a similar angular dependence to the more familiar generalized oscillator strength, it is not identical. For high resolution microanalysis we will primarily be concerned with inner shell excitations, for which a hydrogenic model has been developed by Maslen and Rossouw [22-26]. In this model, an inelastic excitation with scattering vector q for the fast electron is accompanied by ejection of a secondary electron with wavevector κ . As the secondary electron is not detected, κ must be integrated coherently over all secondary electron emission directions, and the matrix product $\rho_{no}(q') \rho_{no}(q' + K)$ replaced with

$$\int \rho_{no}(q', \kappa) \rho_{no}^*(q' + K, \kappa) d\kappa . \quad (13)$$

Convenient analytical expressions have been given for these integrals [27,28]. Fig. 4. shows object functions calculated using the hydrogenic model assuming collection of all the scattering for edge energies appropriate for the Si-K loss (1840eV), Si-L loss (100eV) and the Co-L loss (786eV). Although the hydrogenic model was developed for K shells, we believe it to be the only model for inner shell excitations available at this time, and is likely to provide a more accurate description than simple estimates based on characteristic scattering angles [29-31]. The full width half maxima are 0.24 Å for the Si K loss, 4.9 Å for the Si L loss and 2.0 Å for the Co L loss. Clearly one could not expect atomic resolution from the Si L loss, although it should be possible with the other two edges. Note the sensitive dependence of the object function width on atomic number, the Co L object function being almost an order of magnitude wider than that of the Si K edge, with less than a factor of two difference in energy.

The corresponding functions in reciprocal space, $f'(K)$, are shown in Fig. 5. Clearly, for our detection angle of 30 mrad, almost all of the scattering is indeed detected for the Co-L and Si-L edges, so that the conditions for incoherent imaging are fulfilled. This is not the case for the Si K loss, but provided the detector angle is much larger than the probe forming aperture, incoherent imaging still applies, and the real space object function is now given by restricting the integration in Eqn. 11. to the collector aperture

$$O(R) = \frac{1}{(2\pi)^2} \int_{\text{detector}} f'(K) e^{iK \cdot R} dK . \quad (11)$$

The effect of a finite collector aperture on the object function for the Si K loss is shown in Fig. 6. Clearly, provided we ensure that our collection angle is significantly greater than the objective aperture (semiangle ~ 10 mrad), then the use of a finite collector aperture will not cause a loss in resolution.

Note that all object functions show significant tails in real space, unlike the Gaussian forms often assumed. This will result in a non-local component to the image. It would be expected that this component could be removed by the use of an annular collector aperture to restrict low angle scattering from entering the spectrometer. Fig. 7. shows the significant resolution improvement to be expected with the Si L loss, a FWHM of 3.6 Å and 2.8 Å being achieved for an annular collection apertures of radius 2-30 mrad and 4-30 mrad respectively, which may be sufficient to allow atomic resolution to be achieved in Si. Larger inner collection angles cause substantial loss in intensity.

5. Atomic Resolution Spectroscopy at a $\text{CoSi}_2/\text{Si}(111)$ Interface

The $\text{CoSi}_2/\text{Si}(111)$ epitaxial interface is an ideal test specimen for demonstrating atomic resolution since it can be grown atomically flat and abrupt as seen in the Z-contrast image in Fig 8. This specimen was grown by high dose ion implantation and annealing, and shows an unexpected interface structure in that the last plane of the Si is seen to be in a twinned orientation [32]. The image was taken with the 100 kV STEM, and from the profiles of Fig. 2a would not be expected to resolve the $1.36\text{\AA}$ separation of the dumbbells, but instead each is imaged as a single elongated feature. In the CoSi_2 , the image is dominated by the high scattering power of the Co columns, and the Si columns are unresolved, forming a high background intensity to the image. Note that, due to the twin, the separation of the last plane of the Si and the first plane of the CoSi_2 is only $2.7\text{\AA}$.

Given the object function shown in Fig. 4. we would anticipate achieving atomic resolution using the Co L-loss. The spectra shown in Fig. 9. were obtained by centering a line scan over each plane in turn using the Z-contrast signal displayed on an oscilloscope, then acquiring a Co-L edge for 5 seconds [33]. In this way, the possibility of beam damage is minimised and any specimen drift is visible by observing the line scan. The use of a large collection aperture does degrade the energy resolution of the fine structure, the width of the L_3 edge being about 5 eV. Thus precision in spatial information must be traded for precision in spectral detail due to averaging over a large range of momentum transfers. It is quite apparent that the Co edge drops dramatically in moving a single atomic plane across the interface.

The specimen thickness in this region was estimated at 100-200Å, so that no significant beam broadening should occur in the Si, but some is to be expected for the heavier Co columns. A quantitative profile was obtained by least squares fitting to reference spectra taken several units cells from the interface, and resulted in the profile shown in Fig. 10. The inelastic intensity drops from 86% to 7% in a distance of 2.7Å, demonstrating that atomic resolution is indeed achieved. Also shown is the expected profile for a 2.2Å probe and a perfectly localized inelastic scattering potential. The observed profile is clearly somewhat broadened as would be expected from the significant width of the object function. The asymmetry in the experimental profile is consistent with the anticipated onset of beam broadening in the silicide.

6. Conclusions

Microanalysis at atomic or near-atomic resolution is now experimentally feasible [33-36], opening up new areas of materials science to fundamental study, such as the mechanisms of segregation and embrittlement. The simultaneous acquisition of imaging and spectroscopic data is particularly powerful, due to the complementary nature of their information. For example, in a recent study of tilt grain boundaries in SrTiO₃, Z-contrast imaging and EELS were combined to deduce structure models directly from the experimental data [37-39]. Heavy columns were located by a maximum entropy analysis of the Z-contrast image to an accuracy of ~0.2 Å, while the oxygen coordination was determined from the EELS fine structure. The importance of this capability was the finding of an unexpected structure, involving two half columns (or equivalently a single zig-zag column) at the core of each grain boundary dislocation. Such cores appeared in all structural

units found, and appear to be a very effective way to avoid self-ion repulsion in ionic materials. Also, such columns represent attractive sites for the segregation of dopant atoms, which can now be investigated experimentally by these techniques. Furthermore, the model structures found in this way represent the perfect starting point for theoretical studies, by *ab initio* methods for example. Both theory and experiment may now study the same systems in parallel.

In the future we can also expect to see X-ray analysis carried out at atomic resolution, adding complementary information on elements that are difficult to detect by EELS. We can also expect these techniques to be used with the 300 kV STEM, its smaller probe allowing smaller spaced columns to be illuminated individually, in intermetallics for example. Although the width of the EELS object function increases at the higher accelerating voltage, it does so slower than the probe size decreases, so that real advantages are to be anticipated for materials analysis.

7. Acknowledgments

We are grateful to J. T. Luck, T. C. Estes and S. L. Carney for technical assistance. This research was sponsored by the Division of Materials Sciences, U.S. Department of Energy, under contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

Figure Captions

1. Probe intensity profiles as a function of defocus for (a) the 100 kV HB 501 UX and (b) the 300 kV HB 603 at Oak Ridge National Laboratory.
2. Schematic indicating the columnar channelling of a coherent probe in a zone axis crystal.
3. Schematic indicating the reduction in beam broadening as a result of columnar channelling.
4. Object functions for the Si K, Si-L and Co-L losses, calculated on the hydrogenic model for a 50 eV ejected secondary electron and assuming all scattered electrons are collected. Plots have been normalised to their peak intensity to emphasize differences in width. Note the object function for Si-L is negative at $r=0$.
5. Corresponding object functions in reciprocal space.
6. Reduction and broadening of the Si-K object function due to the use of a finite collection aperture.
7. Enhanced spatial resolution expected with the Si-L loss using an annular collection aperture. Note the significant enhancement of the negative central part of the object function, which will cause loss in signal intensity.
8. Z-contrast image of an epitaxial $\text{CoSi}_2/\text{Si}(111)$ interface showing the last layer of Si to be in a twinned orientation.
9. Co L spectra recorded plane by plane across the interface shown in Fig. 8.
10. Profile showing integrated Co edge intensity plane by plane across the interface (crosses). The sharp drop in Co intensity on moving a single plane across the interface demonstrates atomic resolution. Circles show the theoretical profile for a localized object function.

References

1. L. M. Brown, *Nature* **366** (1993) 721,
2. A. V. Crewe, J. Wall and J. Langmore, *Science* **168**, (1970) 1338
3. A. V. Crewe and J. Wall, *J. Molec. Biol.* **48** (1970) 375
4. A. V. Crewe, J. P. Langmore, and M. S. Isaacson, in: *Physical Aspects of Electron Microscopy and Microbeam Analysis*, eds. B. M. Siegel and D. R. Beaman (Wiley and Sons, New York, 1975), p. 47.
5. O. Scherzer, *J. Appl. Phys.* **20** (1949) 20.
6. A. V. Crewe, *Rep. Prog. Phys.* **43** (1980) 621.
7. A. V. Crewe and D. B. Salzman, *Ultramicroscopy* **9** (1982) 373.
8. C. Colliex and C. Mory, in *Quantitative Electron Microscopy*, eds. J. N. Chapman and A. J. Craven (Scottish Universities Summer School in Physics, Edinburgh, Scotland, 1984) p. 149.
9. S. J. Pennycook, D. E. Jesson and N. D. Browning, *Nucl Instr and Meth* in press (1995)
10. R. H. Ritchie and A. Howie, *Phil Mag A* **58**, (1988) 753
11. D. E. Jesson and S. J. Pennycook, *Proc RSoc Lond* **A441**, (1993) 261
12. D. E. Jesson and S. J. Pennycook, *Proc RSoc Lond* in press (1995)
13. D. McMullan, J. M. Rodenburg, Y. Murooka, and A. J. McGibbon, *Institute of Physics, Conference Series*, **98**, (1990) 55.
14. N. D. Browning, and S. J. Pennycook, *Microbeam Analysis* **2**, (1993) 81.
15. R. D. Leapman, in *Transmission Electron Energy Loss Spectrometry in Materials Science*, eds. M. M. Disko, C. C. Ahn and B. Fultz, The Minerals, Metals & Materials Society, (1992) p. 47.
16. J. Fertig and H. Rose, *Ultramicroscopy* **2** (1977) 269.
17. R. F. Loane, E. J. Kirkland, and J. Silcox, *Acta Cryst.* **A44** (1988) 912.

I_a

a

b

a

b