Crossover Gene Selection by Spatial Location

Dr. David M. Cherba
Computer Science Department
Michigan State University
3105 Engineering Building
East Lansing, MI 48823 USA
cherbada@cse.msu.edu

Dr. William Punch
Computer Science Department
Michigan State University
3105 Engineering Building
East Lansing, MI 48823 USA
punch@cse.msu.edu

ABSTRACT

Spatial based gene selection for division of chromosomes used by crossover operators is proposed for three-dimensional problems. This spatial selection is shown to preserve more genetic material and reduce the disruptive effects of crossover. The disruptive effects of crossover can be quantified by counting the destruction of subgraphs that represent strong linkages between genes. The spatial operator is compared to simple crossover on a practical class of molecular clustering searches. This comparison shows that the spatial crossover significantly out performs simple crossover. Consistent good performance for spatial crossover is demonstrated on the molecular cluster conformation problem [9].

Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Geometrical problems and computations; G.1.6 [**Optimization**]: Global optimization; J.2 [**PHYSICAL SCIENCES AND ENGINEERING**]: Physics

General Terms

Algorithm Crossover

Keywords

Genetic Algorithm, Molecular Conformation

1. INTRODUCTION

The purpose of a crossover operator is to recombine good genetic material existing in the population into more competitive individuals. The building block concept relies on combination of smaller building blocks into larger ones to make progress towards a solution [4]. Numerous studies have been conducted comparing one type of crossover operator to another. In general the results have shown that for any given problem, one crossover operator or a specific combination of operators, might provide better performance than another [13, 14, 8].

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.

The proposed spatial crossover works primarily in phenotype space and is limited to a class of problems where this phenotype crossover has a strong relationship to the fitness function value. The reason is that the co-location relationships between the genes are important and the methods that work well for any given problem preserve the existing good relationships. The conceptual comparison of this crossover operation on phenotypical expression to more indirect encoding of genes is based on the final goal of all crossover operators to preserve and combine genetic material regardless of encoding.

The use of gene linkage is an issue with genetic representation, operators and Genetic Algorithm (GA) control. The various methods for determining linkage between genes can be complex and often limited to simple groupings of genes. Rather than solve those issues we choose to avoid them by introducing an operator that uses phenotypical expression of genes to perform the crossover operation in phenotype space. This maintains the phenotype/physical linkages without regard to genotype linkages.

Seo *et al.* [12] explores the topological linkages of genes and lists many of the current linkage methods. In this work, the introduction of non-linear division takes an arrangement of genes expressed in the phenotypical space and does the crossover division using an arbitrary set of boundaries in that space. Non-linear in this context is any method that does not work on a linear arrangement of genes on a chromosome. There is still a disconnection between the arrangement of genes and the boundary selection methods. Spatial selection of genes for crossover preserves existing good genetic material in such a way that it can be combined into more competitive individuals. The connection between the genes that have strong relationships are maintained.

Crossover operators can be characterized by the number of points at which a chromosome is divided and the methods by which the points are chosen. The taxonomy is complicated by the ways in which the gene values are combined. In real value coding there are a number of methods for numerical computation combining the values of gene from both parents or even multiple parents [6, 11, 1] . For binary coding most crossover methods are discrete in nature, choosing a gene from one parent or the other to include in the offspring. All the discrete methods associated with binary coding of genes can also be applied to real valued genes. In the spatial crossover the method of selecting which genes are contributed to the offspring is based on the phenotypical expression of the gene in relation to other genes expressions not random or sequential selection.

Two issues with crossover operators include positional bias [3] and disruption. The impact of disruption was analyzed for single point crossover by Holland with the introduction of schemata [7]. Multi-point crossover was the focus of a work by Spear [13]. All possible sets of schemata are equal and the impact of multiple division points will be more likely to disrupt larger schema whose genes are spread across the chromosome. Gene to gene relationships have a major impact on the objective function and how the crossover operators preserve or disrupt those relationships often determines if the problem can be solved in a computational effective manner.

Many of the advanced methods in genetic algorithms focus on learning the effects that these relationships have and altering operations to preserve the good relationships. Linkage learning has been used to improve the performance of genetic algorithms. If a gene belongs to only one relationship then the use of crossover operator that tends to preserve a single relationship will produce better performance. The problem is that for many genes there are multiple significant relationships to other genes.

The use of real-valued genes has many practical advantages over binary encoding. However, it can be argued that all real-value encoding can be converted to binary. When real-valued genes are used, the control of mutation magnitudes and possible numeric combinations are more straight forward to conduct. Real-valued vectors for locations facilitates the use of standard vector operators like dot and cross products in the fitness functions and selection processes.

Many genetic algorithm applications try to solve problems that model physical entities like mechanical designs, atom locations, and flow quantities. The natural representation of candidate solutions are often based on real-valued genes. Further, the objective function can be greatly affected by the location of each gene in space relative to other genes. Regardless of how the actual gene is encoded the expression of these genes must eventually translate to a location of a point in space. The value produced by the objective function will be greatly influenced by the distance between points.

One such class of problems is the molecular conformation problem. In molecular conformation problems the value of the objective function is more strongly influenced by close point distances. If all the points could be arranged such that a simple linear placement of each gene on a chromosome maintains close proximity of points with the greatest effect on each other, then it would be possible to select from the current collection of genotype space crossover methods.

It is rarely possible to create such an arrangement of genes for three-dimensional points. Some attempts to arrange the genes in grids, lattices, and more complex organizations with flexible boundaries have demonstrated good results but they require consistent mapping from the physical location in space to the organization of the gene arrangement [10, 12]. This has limited the application of the various gene arrangements to a specific sets of problems like Very Large Scale Integration (VLSI). The molecular conformation problem is an example of three-dimensional mapping problem. This general problem does not lend itself to using a consistent uniform lattice arrangement. In some cases such as hydrophobic protein analysis, simplified unit dimensional grids are used. Methods for moving the gene location on the chromosome have also been used to try and group the closely linked genes.

A significant result for genetic algorithms applied to molecular conformation occurred when Deaven and Ho [2] were able to find the structure of a variety of fullerene molecules including the Bucky ball using a genetic algorithm. The introduction of a spatial based crossover method was a key difference in their GA algorithm. The spatial crossover operator is applied to the phenotypical expression of the genes, preserving the relationship between genes that are close in the phenotypical space. While the algorithm also used diversity control and local search, the implementation of the crossover was critical to the overall operation of the algorithm. In addition, the use of spatial crossover has been adopted by Hartke [5]. To date there has been no theoretical analysis of why this crossover operator provides such good performance.

This work will provide insight on why this spatial crossover method is well suited for this class of problems. A variety of spatial gene selection processes are possible using different dividing elements to partition the genes into sets. Only a simple process using a single dividing element consisting of a three-dimensional plane will be analyzed because of it's application to the molecular cluster conformation problem. In general, the space can be divided by many elements.

Viewing the phenotypical expression of the genes as locations in space will demonstrate the disruptive effects that crossover can have on the spatial relationships. It is possible to quantify the disruption by assignment of edges to the relationships followed by examination of the effect of spatial and simple crossover on those edges. The nodes that are close to each other are considered to be more strongly linked and this is represented as edges between the nodes. Counting the subgraphs that are disrupted by the crossover operators demonstrates the impact that each operator has on the linkages between the genes. These subgraphs form a subset of the total schemata for all the genes. The last section will describe a simple comparison experiment between two-point crossover and the spatial crossover applied to the molecular conformation problem.

2. PROBLEM

The problem is to explore is why this spatial crossover works so well for the molecular conformation search. Looking at the basic concept of schemata, some observations need to be made. For a schema, the gene location on the chromosome is not critical but its location is normally fixed. A specific gene is normally rooted at a specific location on the chromosome and does not change location because keeping track of genes movement on the chromosome would complicate the process. All that should be important is that high fitness parents contain collections of genes that contribute to the high fitness. Eventually, by combining high fitness genes, larger building blocks with high fitness will be created. This is likely not the case where genes locations are fixed. If the same gene location x on each parent contained different genetic material (i.e. like a different but needed 3-D point) then recombination will not produce the larger building block incorporating points from both parents using discrete crossover on fixed location genotypical encoding. Further, if the genes were significantly different, then only large scale mutations would convert another gene on the chromosome to the correct value. The only hope would be that a mutation on another gene of either parent would eventual create the correct material.

Even though the fitness of both parents are high the material will never get a chance to combine by crossover. The concept of combining small genetic building blocks into larger building blocks would be limited in this environment.

Given randomly generated genes that represent points in space, it is possible that the same gene location on many members of the population will contain different but needed genetic material. Furthermore there will be genes in different locations that will contain duplicate (or nearly so) material. This second issue can cause more trouble than the first for some problems like the molecular conformation problem. Atoms too close together have problems with energy models. In real physics they cause highly exothermic reactions instead of computer overflows. The problem is to find a selection method that allows for combining good genetic material into larger building blocks and that does not destroy the current good building blocks. Further, it must allow for many overlapping building blocks.

One way to quantify the building blocks in these type of spatial problems is to count membership in subgraphs. Because each point may be part of several subgraphs it is necessary to count all the subgraphs memberships rather than just points in a single building block. For example, a point that is part of only one building block or subgraph is less significant than one that is part of several building blocks or subgraphs. This is especially true in the molecular problem where each atom position relative to four or five other atoms effects the fitness of a given building block.

3. SPATIAL GENE SELECTION

Standard crossover operators use a variety of methods to determine which genes to select for recombination into an offspring. These methods normally have one or more random choices that identify the cutting points along the chromosome. For genes that represent phenotypical locations in space there is no expectation that adjacent genes will be in the same region. Further, random selection of genes from each parent will not do any better at collecting good building blocks into larger ones for locations that need to be in a local region of space for high fitness. Mapping points into space and then dividing the space will preserve the localized relationships between points. Relationships along the boundary of dividing element will still be disrupted or possibly create new needed material. This allows the division of the points into cohesive set by any number of methods. The selection method that divides each parent into two spatially coherent parts starts with the center of mass and a dividing elements.

For the simple two-dimension case shown in Figure 1 the space is divided by a single line. The center of mass can adjusted so that equal numbers of points are on either side of the line. For odd numbers of points the the balance is adjusted to be as close to equal as possible. The orientation of the line is chosen at random. Given that the locations can reside anywhere on the chromosome, determination of which partition to place each node is accomplished with the dot product between a chosen normal vector to the line through the center of mass and the vector from the center of mass to the location. Positive dot products are in one set and negative products are in the other set. The same random line orientation is used on both parents.

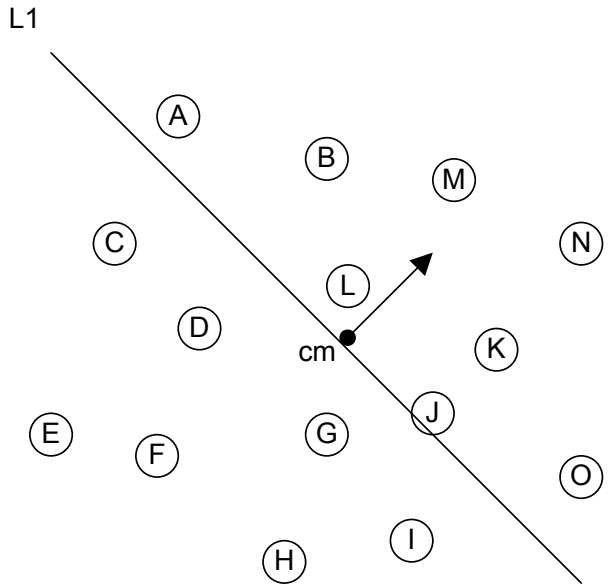

Figure 1: Simple Spatial selection process

However, the center of mass will vary by some small amount. In the Figure 1 the center of mass is at the origin of the vector shown normal to line L1.

In general, the locations can be in X dimensional hyperspace and the dividing element can be a X-1 dimensional hyperplane. The vector normal to the plane is used along with the vector from the center of mass to the locations. For applications that are N-dimensional the same basic process can be used. Assume N location vectors, each of dimensionality M. Each of the N vectors represent a point in space $p_1 \ldots p_N$. In Equations 1, 2 and 3 the suffix's n and p indicate which side of the dividing plane the location is on. The V_n is the normal to the hyperplane.

$$\mathbf{P} = \mathbf{P}_p \cup \mathbf{P}_n \tag{1}$$

$$\forall\, p_i \ \in \mathbf{P} \mid p_i \circ V_n \geq 0\ \mathbf{P}_p = \mathbf{P}_p \cup p_i \tag{2}$$

$$\forall\, p_i \ \in \mathbf{P} \mid p_i \circ V_n < 0\ \mathbf{P}_n = \mathbf{P}_n \cup p_i \tag{3}$$

Although this is shown for single division there is no reason that multiple divisions can not be used, making the partitions small in a way is similar to genetic multi-point crossover. The downside to this is the disruption of the subgraphs will be increased but the ability to combine smaller pieces will be greater.

The division process is straight forward, however it is still possible for two parents to have locations near the dividing element that violate a minimum distance rule when combined. This minimum distance rule is intended to prevent poor chromosomes with locations too close together. When the parts from each parent are combined an adjustment process moves the locations in the most direct way to satisfy the minimum distance constraint. This process occasionally fails and the offspring is thrown away.

4. SUBGRAPH COUNTING

Analysis for the crossover effect on this class of problems is based on counting membership in subgraphs. A subgraph is defined as N locations connected by $N-1$ edges. The order of each subgraph is labeled by its number of locations. In the examples provided, all the subgraphs are counted. However for this class of problem those of orders $\{2,3,4,5\}$ are the most significant because they will have the greatest influence on the fitness function typically associated with this class of three-dimensional problem. Only locations that are close together would be counted because the assigned edges are limited in length and therefore connecting only close locations. Limiting the length of edges corresponds to the relationships that close locations produce a greater effect on the fitness values than locations far apart.

Two subgraphs are considered equal if the set of nodes and edges are the same and only one of them is counted. Two subgraphs are consider different if the set of nodes is the same but the edges are different and both are counted.

The number of cut points used in classic crossover operators will increase the likelihood of disrupting good schema [13]. In the case of spatial crossover operators, this is also true. The equivalence of a single cutting plane for spatial crossover is similar to two-point classic crossover for a linear arrangement of genes where the two points must divide the chromosome into equal halves. Given the definition of subgraphs as building blocks or a type of schema, it is instructive to compare the disruption of those subgraphs by the two crossover methods.

Three examples will be presented. First is the simple cube that can be easily visualized and the subgraphs counted by hand. The second example is a molecule from the table of known low potential energy configurations using eleven atoms. The final example is a Bucky ball that was the target of a GA search for molecular conformation from an experimental data set [9].

4.1 Simple Cube

A cube as shown in Figure 2 with labeled nodes. In this example the goal is to enumerate all the (building (blocks / subgraphs)) associated with the cube to demonstrate the concept of subgraph counting for spatial applications. The N2 building blocks consist of the edges in the cube. For the N3 building block, each vertex is at the center of three unique N3's, so there are a total of 24. It should be noted that in normal schema, each combination of items would be valid. The set representation for the combination of eight nodes taken two at a time would have a total of 28 members. Overall the total count is reduced from 512 to 158 by the restrictions of the connected graph. For this concept the subgraph has to be connected by a unique set of edges. This eliminates many of the possible building blocks. The full table of building blocks enumerated by size is shown in Table 1. This table also shows the number of building blocks that are destroyed by a cutting plane and two point crossover. All the building blocks above N4 are destroyed. Even for this simple example of the cube, the disruption of the small subgraphs is significantly reduced by the spatial crossover compared to two point crossover. The number of disrupted subgraphs for the two-point crossover was found by numeric average then rounding down using one hundred trial selections. It was possible to disrupt all subgraphs with a bipartite numbering of the nodes and this occurred a significant number of times in the one hundred trials.

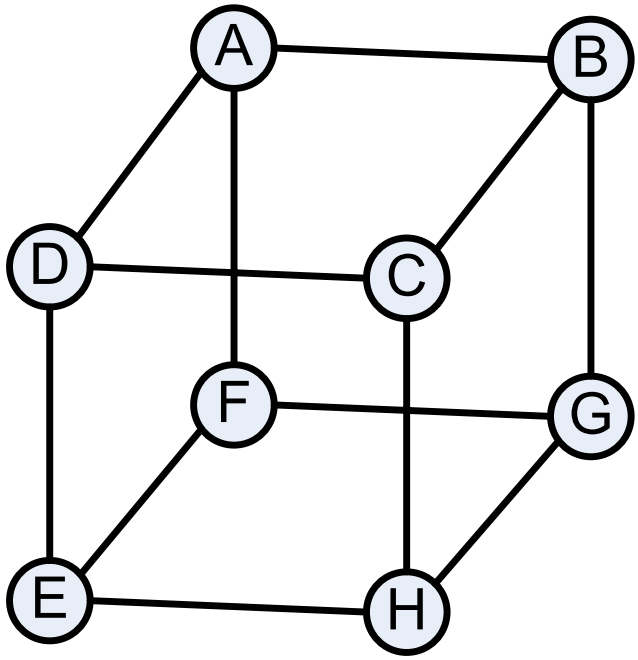

Figure 2: Cube schema R_3

order	Number	Broken by Plane	2 Point
N2	12	4	10
N3	24	16	20
N4	38	36	35
N5	48	48	48
N6	28	28	28
N7	8	8	8
N8	1	1	1

Table 1: Building blocks / subgraphs for cube

Even in this very simple example the cutting plane destroys fewer subgraphs than the random selection associated with the two point crossover. Therefore if the subgraphs are representative of good building blocks, then using spatial crossover will preserve more good building blocks than two-point crossover.

4.2 R regular graphs

The Bucky ball used in the original molecular conformation work is an R-regular graph where all the vertex are order three [9]. The problem is that counting the total number of subgraphs for a sixty node graph would not provide as much information as estimating the number of subgraphs for a limited number of nodes. In this case the limit is set at five nodes. The Table 2 shows an estimate for the number of subgraphs and the disruptive effect of the two crossover methods. The two-point crossover disruption of subgraphs was calculated by a conservative probability estimate.

In this estimate the chance of two adjacent being in the same half was assumed to be .5. The total number of subgraphs for each order was produced by counting the unique number of subgraphs for a single pair of nodes and using symmetry to find the total.

order	Number	Broken by Plane	2 point
N2	90	14	45
N3	180	56	135
N4	420	168	368
N5	1080	672	1012

Table 2: Bucky building blocks

From this table, it is clear that, especially for the Bucky ball, the number of disrupted subgraphs is much smaller for plane division than from two-point crossover. The reduction in the disruption is on the order of 3 : 1 for the smaller subgraph sizes.

4.3 N11 Lennard-Jones Molecule

Now that the general method for counting subgraphs for large molecules modeled as R regular graphs and a simple molecule arranged as a cube have been examined, it is appropriate to explore a moderate-sized molecule. The N11 Lennard-Jones molecule configuration is examined. For this molecule the size of the subgraph will be described by the number of nodes it contains and then the total number of subgraphs of that size. In this example, the size of each subgraph is shown in the first column. The number of total subgraphs at each size is shown in the second column. The third and fourth columns show the number of subgraphs that are broken by a plane division and two-point crossover respectively. In the Table 3, the number of smaller subgraphs that are disrupted is significantly less with the plane division than with two-point crossover. The number of subgraphs with size less than five atoms or nodes shows a reduced number of disruptions. As the size of the subgraph grows, then just about any division in the chromosome is likely to disrupt the subgraph.

These three examples from a variety of arrangements show a consistent qualitative measure that indicates spatial crossover selection by means of a dividing plane disrupt fewer subgraphs than two point crossover.

order	Number	Broken by Plane	2 Point
N2	31	12	28
N3	91	65	86
N4	215	195	210
N5	374	367	369
N6	438	437	435
N7	328	328	328
N8	165	165	165
N9	55	55	55
N10	11	11	11
N11	1	1	1

Table 3: Building blocks for ideal N11 LJ

N20	Two-Point	Spatial
Failures	54	0
Average	491.3	35.65

Table 4: Solution comparison of two-point and spatial crossover for N20 Lennard-Jones molecules

N38	Two-Point	Spatial
Failures	100	4
Average	NA	306.

Table 5: Solution comparison of two-point and spatial crossover for N38 Lennard-Jones molecules

5. EXPERIMENTAL RESULT

To formulate a simple comparison between spatial crossover and the two-point crossover, a substitution of one method for another on a working problem was selected. The molecular conformation problem using only distance data was chosen as the example problem. The experimental data or ideal data set consist of $(n-1)n/2$ distances between atoms and is the target of the objective function. The genetic algorithm is trying to configure molecules in the population so that individuals in the population will have distances between atoms that match the target set of distances provided by the experimental data or ideal data.

The spatial crossover was able to solve the twenty atom Lennard-Jones configuration 100% of the time in a few hundred generations. Replacing the spatial crossover with two-point crossover and running the same trial produced the results shown in Table 4. The only parameter that was changed was the method of crossover. The population size, operator probabilities, and iteration counts were all kept the same. The same local search was also used with the unchanged parameters. This local search is based entirely on the distance between atoms. It applies incremental corrections to the location of atoms based on vector errors calculated by the differences between the target distances and distances between pairs of atoms in the individual member of the GA population. When spatial crossover was replaced with the two-point crossover method, the ability to find solutions was severely degraded. In the Lennard-Jones twenty atom molecule, the algorithm only found 46 solutions in a 100-trial series. The average generations to find a solution for successful runs increased by over a factor of ten.

Another set of experiments was conducted on the N38 or thirty-eight atom Lennard-Jones molecule and the results were worse. In this case, 100% of the trials failed to find a solution in the maximum allotted number of 4000 generations. Expecting that several other GA parameters beside the crossover were having a significant negative impact, a series of parameter sweeps were conducted to discover better parameters. In all cases, there was either no change or further degradation as measured by the best fitness achieved during the run. Based on this, only the N20 and N38 series of runs using the expected optimum parameter settings are reported. Table 5 shows the summary of the results for the N38 Lennard-Jones molecule.

6. CONCLUSION

It is clear from counting subgraphs that spatial crossover will be less disruptive than two-point crossover for problems where the relationship between phenotypical expression of genes is based on location and distance. The simple example for the cube and also the molecule models shows that for two-point crossover it would be very difficult to keep good building blocks intact during crossover. Spatial crossover shows that it will preserve more building blocks and therefore allow solutions to evolve. The counting of subgraphs for even large molecules can be estimated using combinatorial math with equivalent sets removed for a modest order of subgraphs less than six nodes. The experimental results clearly shows that for this class of problems that the spatial crossover out performs two-point crossover as the size of the problem increases. Several GA researchers have used small molecules on the order of $3-11$ atoms with success. The result for the thirty-eight atom molecule shows the effect that size has on the problem as compared to the twenty atom molecule.

The theory to quantify the performance of spatial crossover as compared to two-point crossover shows that larger molecules will pose significantly greater problems for two-point crossover. The experimental results supports this observation from the theory.

For chromosomes genes that are expressed as locations the spatial crossover has shown superior performance. It is clearly not intended to be an universal solution however, it does offer the potential to be a much needed tool when working with problems that are formulated with distances and locations in space as key components. Many problems can be formulated as N-dimensional spatial problems like the molecular conformation search.

7. REFERENCES

[1] J. Arabas, J. J. Mulawka, and J. Pokrasniewicz. A new class of the crossover operators for the numerical optimization. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 42–48. Morgan Kaufmann Publishers Inc., 1995.

[2] D. Deaven and K. Ho. Molecular geometry optimization with a genetic algorithm. *Phy. Rev. Let.*, 75:288–291, 1995.

[3] L. J. Eshelman, K. E. Mathias, and J. D. Schaffer. Crossover operator biases: Exploiting the population distribution. *Proc. of the 7th ICGA*, pages 354–361, 1997.

[4] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[5] B. Hartke. *Application of Evolutionary Computation in Chemistry*, volume 110 of *Structure and Bonding*, chapter Application of Evolutionary Algorithms to global cluster geometry optimization, pages 33–53. Springer, Heidelberg, 2004.

[6] F. Herrera, M. Lozano, and A. Sånchez. A taxonomy for the crossover operator for real-coded genetic algorithms. an experimental study. *International Journal of Intelligent Systems*, 18(3):309–338, 2003.

[7] J. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Pres, Ann Arbor, 1975.

[8] I. Hong, A. B. Kahng, and B. R. Moon. Exploiting synergies of multiple crossovers: initial studies. *IEEE International Conference on Evolutionary Computation*, 1(29):245–250, Nov. Dec. 1995.

[9] P. Juhas, D. M. Cherba, P. M. Duxbury, W. F. Punch, and S. J. L. Billinge. Ab initio solid state nano-structure determination. *Nature*, pages 655–658, March 2006.

[10] A. B. Kahng and B. R. Moon. Toward more powerful recombinations. In L. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 96–103, San Francisco, CA, 1995. Morgan Kaufmann.

[11] Z. Michalewicz, G. Nazhiyath, and M. Michalewicz. A note on usefulness of geometrical crossover for numerical optimization problems. *Evolutionary Programming*, pages 305–312, 1996.

[12] D.-i. Seo and B. R. Moon. A survey on chromosomal structures and operators for exploiting topological linkages of genes. In *GECCO*, pages 1357–1368, 2003.

[13] W. M. Spears and K. A. De Jong. An analysis of multi-point crossover. In G. J. E. Rawlins, editor, *Foundations of Genetic Algorithms*, pages 301–315, San Mateo, CA, 1991. Morgan Kaufmann.

[14] H.-S. Yoon and B.-R. Moon. Synergy of multiple crossovers in a genetic algorithm. *IEEE Transactions on Evolutionary Computation*, 6(2):212 – 223, April 2002.