Genotype-Fitness Correlation Analysis for Evolutionary Design of Self-Assembly Wang Tiles

Germán Terrazas and Natalio Krasnogor

Abstract In a previous work we have reported on the evolutionary design optimisation of self-assembling Wang tiles. Apart from the achieved findings [11], nothing has been yet said about the effectiveness by which individuals were evaluated. In particular when the mapping from genotype to phenotype and from this to fitness is an intricate relationship. In this paper we aim to report whether our genetic algorithm, using morphological image analyses as fitness function, is an effective methodology. Thus, we present here fitness distance correlation to measure how effectively the fitness of an individual correlates to its genotypic distance to a known optimum when the genotype-phenotype-fitness mapping is a complex, stochastic and non-linear relationship.

1 Introduction

Self-assembly systems are characterized by inorganic or living entities that achieve global order as the result of local interactions within a particular closed environment. Self-assembly is a key cooperative mechanism in nature. Surprisingly, it has received (relatively) very little attention in computer science. In [11], we defined the *self-assembly Wang tiles system* T_{sys} as a computational model of self-assembly. This system comprises a set of square tiles with labelled edges that randomly move across the Euclidean plane forming aggregates or bouncing off as result of their in-

Germán Terrazas
ASAP Group, School of Computer Science
University of Nottingham, UK
e-mail: gzt@cs.nott.ac.uk

Natalio Krasnogor
ASAP Group, School of Computer Science
University of Nottingham, UK
e-mail: nxk@cs.nott.ac.uk

teraction (see Fig. 1). Cooperativity is an emergent feature of this system where the combination of a certain number of tiles is required to initiate self-assembly [5][9].

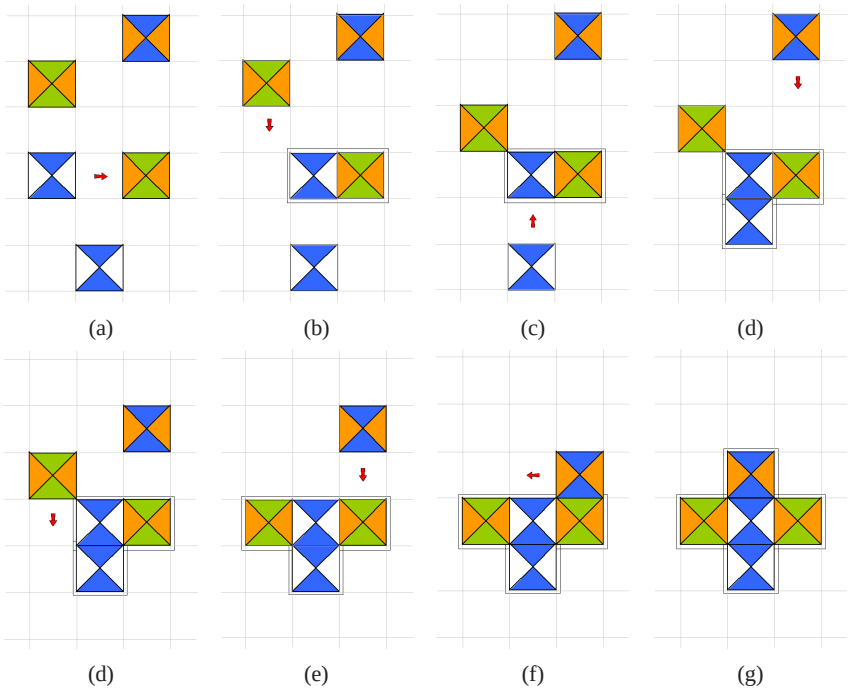

Fig. 1 A step-by-step (a to g) aggregate formed as the result of the interaction between five self-assembly Wang tiles performing random walk across a lattice.

Formally speaking, T_{sys} is defined as follows:

$$\begin{aligned} T_{sys} &= (\mathscr{T}, \Sigma, g, \mathscr{L}, \tau) \\ \mathscr{T} = \{t | t = (c_0, c_1, c_2, c_3)\} &\text{ where } c_0, c_1, c_2, c_3 \in \Sigma \\ g &:: \Sigma^2 \to \mathbb{Z}^+ \cup \{0\} \\ \tau &\in \mathbb{Z}^+ \cup \{0\} \end{aligned} \tag{1}$$

In this system, $\mathscr{T}$ is a finite set of non-empty Wang tile types t defined as a 4-tuple (c_0, c_1, c_2, c_3) which indicates the associated labels at the north, east, south and west edges of a tile, Σ is a set of symbols representing glue type labels, g is called the *glue function*, $\mathscr{L}$ is a lattice and τ is a threshold modelling the kinetic energy of the system. The glue function g is defined to compute the strength associated to a given pair of glue types. The lattice $\mathscr{L}$ is a two-dimensional surface with size $W \times H$ composed by a finite set of interconnected unit squared cells where tiles belonging to $\mathscr{T}$ are randomly located and perform random walks. Thus, when two or more tiles collide, the strength between the glue types at their colliding edges is calculated and

compared to τ. If the resulting strength is bigger than τ, tiles self-assemble forming aggregates, otherwise they bounce off and keep moving.

Finding the appropriate combination of autonomous entities capable of arranging themselves together into a target configuration is a challenging open problem for the design and development of distributed cooperative systems. In [11] we addressed the self-assembly Wang tiles designability problem by means of artificial evolution. Our interest in combining self-assembly Wang tiles with evolutionary algorithms lays on the use of a method for the automated construction of supra-structures that emerge as the result of tiles interaction. In particular, we pursued to answer the following:

Given a collective target configuration, is it possible to automatically design, e.g. with an evolutionary algorithm, the local interactions so they self-assemble into the desired target?

In order to address this question, our research was centred in the use of a genetic algorithm (GA) to evolve a population of self-assembly Wang tile families. Broadly speaking, a self-assembly Wang tile family is a descriptor comprising a set of glue types each of which is associated to one of the four sides of a self-assembly Wang tile. Thus, each tile family is instantiated with equal number of tiles which are randomly located into a simulation environment where they drift and interact one another self-assembling in aggregates until the simulation runs it course. Once the simulation finishes, aggregates are compared for similarity to a user defined (target) structure employing the Minkowski functionals [6] [7]. This assembly assessment returns a numerical representation that is considered as the fitness value (Fitness_i) of each individual. Thus, individuals capable of creating aggregates similar to the specified target are better ranked and become the most likely to survive across generations. This process, together with one-point crossover and bitwise mutation operators, is applied to the entire population and repeated for a certain number of generations. In particular, our experiments comprised four increasingly rich/complex simulation environments: deterministic, probabilistic, deterministic with rotation and probabilistic with rotation. An illustration summarising our approach and its components is shown in Fig. 2.

The achieved results supported our evolutionary design approach as a successful engineering mechanism for the computer-aided design of self-assembly Wang tiles. Early evidence of our research in this topic is available in [10] where we employed a very simple evaluation mechanism composed by a lattice scanner fitness function and later, in [11], where morphological image analyses brought a more accurate and efficient way to collectively assess the assembled aggregates towards the target. Since the mapping from genotype to phenotype and from this to fitness is clearly a complex, stochastic and non-linear relationship, would be possible to analyse the effectiveness of Minkowski functionals as fitness function ? In order to address this question, we first introduce fitness distance correlation. Next we present how this statistical-based protocol is applied to analyse our method together with experiments and results. Finally, conclusions and discussions follow.

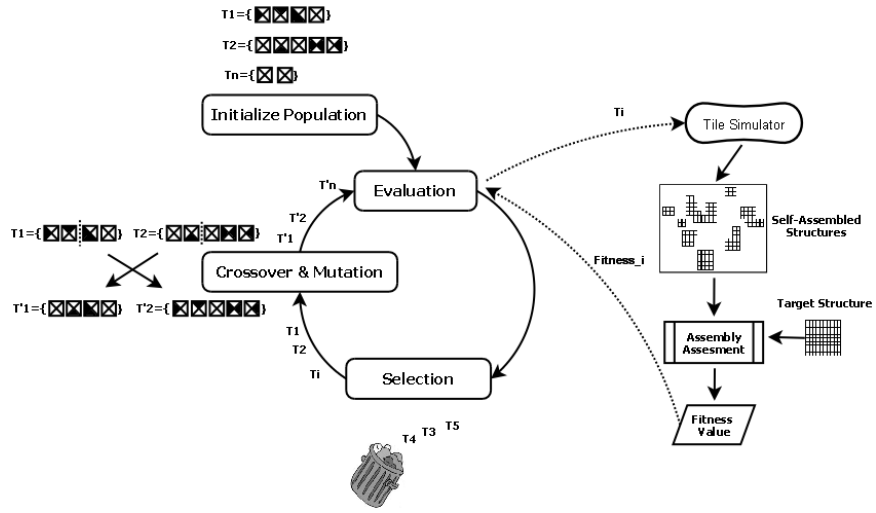

Fig. 2 Evolutionary approach for the evolutionary design optimisation of self-assembly Wang tiles. A population of self-assembling Wang tiles family (genotype) is randomly initialised. After that, each individual is set up into a tiles simulator from where the emerging self-assembled aggregations (phenotypes) are compared against a target structure for similarity. This comparison returns in the fitness of the individual. Later on, the application of genetic operators follows where the best ranked individuals are likely to pass throughout selection, crossover and mutation stages.

2 The Genotype-Fitness Assessment Protocol

The evolutionary design of self-assembly Wang tiles is characterised as a problem in which the mapping from genotype to phenotype and then from phenotype to fitness is a highly complex, non-linear and in some cases stochastic relationship. It is non-linear because different genotypes (tile sets) with small differences may lead to widely diverging phenotypes. While the same genotype, due to random effects, might produce a variety of end-products. This intricate relationship (see Fig. 3) makes the assessment of the genotype very difficult since the same (different) fitness value could be assigned to different (the same) genotypes. Hence, in order to analyse the efficiency by which individuals were evolved, we employed Fitness Distance Correlation (FDC) [3] [2] to measure how effectively the fitness of an individual correlates to its genotypic distance to a known optimum.

FDC is a summary statistic that performs a correlation analysis in terms of a known optimum and samples taken from the search space, predicting whether a GA will be effective for solving an optimisation problem. Thus, when facing a minimisation (maximisation) problem, a large positive (negative) correlation indicates that a GA may successfully treat the problem or that the problem is straightforward, whereas a large negative (positive) value suggests that employing a GA may not be effective or that the problem is misleading. However, a correlation around zero, i.e.

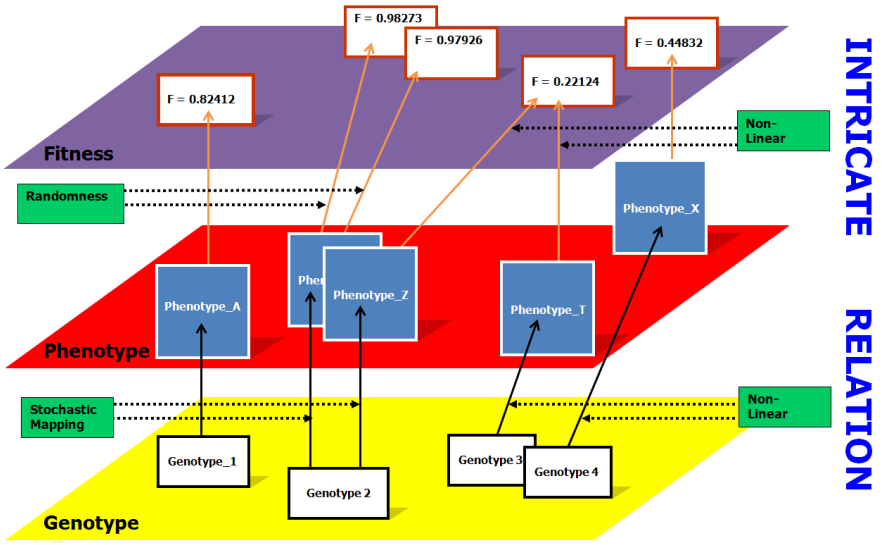

Fig. 3 The highly complex, non-linear and stochastic relationship taking place across the mapping from genotype to phenotype and then from phenotype to fitness.

$-0.15 \leq \text{FDC} \leq 0.15$, would suggest that more nuisances, perhaps including scatter plots analyses, of the fitness versus distance to the optimum should be done and, in general the problem is categorized as difficult. The formula for computing the FDC is shown in Equation 2, where n is the number of samples, f_i is the fitness of sample i with distance to the known optimum d_i, $\overline{f}$ and S_F are the mean and standard deviation of the fitness values, and $\overline{d}$ and S_D are the mean and standard deviation of the distances.

$$
\begin{aligned}
FDC &= \frac{(1/n)\sum_{i=1}^{n}(f_i-\overline{f})(d_i-\overline{d})}{S_F S_D} \\
\overline{f} &= \frac{1}{n}\sum_{i=1}^{n} f_i \\
\overline{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
S_F &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(f_i-\overline{f})^2} \\
S_D &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\overline{d})^2}
\end{aligned}
\tag{2}
$$

A study focused on whether FDC predicts the GA behaviour, and whether it detects differences in encoding and representation for a number of well-studied min-

imisation problems has been given in [3]. When predicting the GA behaviour, the FDC confirmed that Deb & Goldberg's 6-bit fully deceptive function and Whitley's 4-bit (F2 and F3) fully deceptive functions are indeed misleading since the correlation values were 0.30, 0.51 and 0.36 respectively, and the fitnesses tended to increase with the distance from the global optimum. In addition, FDC also confirmed that problems like Ackley's One Max, Two Max, Porcupine and NK landscape problems for $K \leq 3$ are easy since the correlation values resulted in -0.83, -0.55, -0.88 and -0.35 respectively. Nevertheless, the FDC indicated that NK(12,11) landscape, Walsh polynomials on 32 bits with 32 terms each of order 8, Royal Road functions R1 and R2, as well as some of the De Jong's functions like F2(12) are difficult since the resulting correlation values were close to 0.0. When the differences in encoding and representation were considered, experiments using Gray and binary coding led to the conclusions that the superiority depends on the number of bytes used to encode the numeric values. For instance, De Jong's F2 with binary coding is likely to make the search easier than with Gray coding when using 8 bits. In contrast, the correlation value of F12 indicated that Gray coding works better than binary when using 12 or 24 bits. Despite its successful application on a wide benchmarking set of problems, FDC is still not considered to be a very accurate predictor in some other problems. For instance, a case where FDC failed as a difficulty predictor has been presented when studying a GA maximising *Ridge functions* [8].

In summary, although FDC cannot be expected to be a perfect predictor of performance, previous work reported in [4] [1] [13] [12] [14] suggests that it is indeed a good *indicator* of performance. Our goal is then to assess how effectively the fitness of an individual correlates to its genotype when using Minkowski functionals as fitness function for the GA presented in [11].

3 Correlating the Self-Assembly Wang Tiles Design

Since different set of tiles may self-assemble in aggregates similar in shape to the target structure, it is of our interest to study here how effectively the fitness of an individual correlates to its genotypic distance to a known optimum (see Fig. 4). In the rest of this section, we carry out FDC analysis in order to study Minkowski functionals effectiveness as fitness function for the evaluation of the achieved self-assembled aggregates.

3.1 Experiments and Results

In order to perform a FDC analysis, we first choose the best individual found among the four simulation environments. In this case, the best individual belongs to the results achieved when using probabilistic criteria and no rotation simulator. Next, a

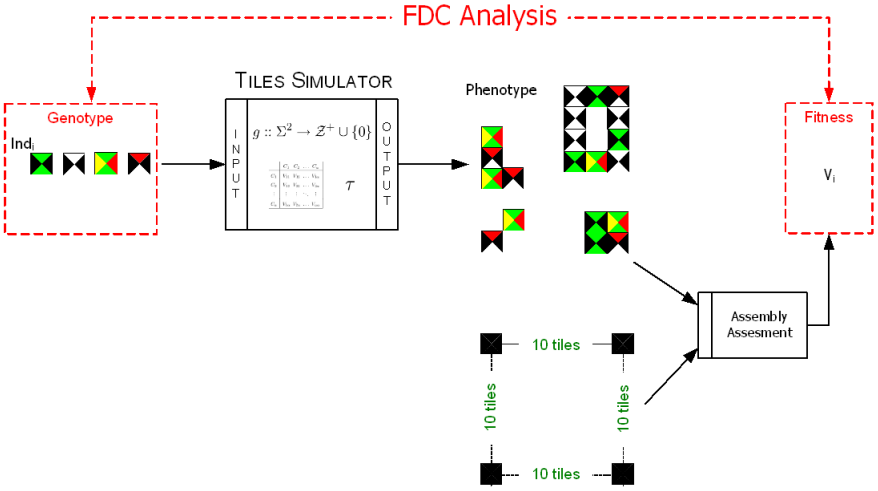

Fig. 4 Diagram of mappings from genotype onto phenotype and from phenotype onto numerical fitness value, and relationship to the Fitness Distance Correlation.

data set comprising 500 individuals at different Hamming distances from the best individual is created. In particular, given two individuals Ind_i and Ind_j of same length, their Hamming distance H is defined as in Equation 3.

$$
\begin{aligned}
H(Ind_i, Ind_j) &= \sum_{k=1}^{n} diff(T_k^i, T_k^j) \\
diff(T_i, T_j) &= \sum_{l=0}^{3} c_l^i \ominus c_l^j \\
c_a \ominus c_b &= \begin{cases} 1 \text{ if } c_a \neq c_b \\ 0 \text{ otherwise} \end{cases} \\
c_a, c_b &\in \Sigma
\end{aligned}
\tag{3}
$$

Thus, this 500 individuals data set comprises all the possible chromosomes at Hamming distance of 1 plus some other randomly generated individuals at greater distance, all of these systematically generated following the pseudocode described in Algorithm 1 where $DuplicateReplacing(T_k, c_l, c_{new})$ duplicates tile T_k replacing glue type c_l with c_{new}, $DuplicateReplacing(Ind_i, T_k, T_{new})$ duplicates individual Ind_i replacing tile T_k with T_{new}, $TileAt(Ind_i, k)$ returns the tile at position k of an individual, and $Replace(T_k, c_l, c_{new})$ replaces glue type c_l in tile T_k with glue type c_{new}.

Algorithm 1 GenerateIndividuals

Require: *Ind* an individual
Ensure: *S* a set of individuals
1. **for all** tiles T_k in Ind_i **do**
2. **for all** glue types c_l in T_k **do**
3. **for all** glue type $c_{new} \in \Sigma$ **do**
4. $T_{new} \leftarrow DuplicateReplacing(T_k, c_l, c_{new})$
5. $Ind_{new} \leftarrow DuplicateReplacing(Ind_i, T_k, T_{new})$
6. $Insert(S, Ind_{new})$
7. **end for**
8. **end for**
9. **end for**
10. **while** $|\, S \,| < 500$ **do**
11. $Ind_{new} \leftarrow Duplicate$(randomly chosen $Ind_i \in S$)
12. $n \leftarrow Random(0, |\, Ind_{new} \,|)$
13. **for all** k to n **do**
14. $T_k \leftarrow TileAt(Ind_{new}, k)$
15. $m \leftarrow Random(0, 3)$
16. **for all** l to m **do**
17. $c_{new} \leftarrow Random(\Sigma \setminus c_l)$
18. $Replace(T_k, c_l, c_{new})$
19. $Insert(S, Ind_{new})$
20. **end for**
21. **end for**
22. **end while**

In total, each of the generated individuals was simulated 5 times giving as a result a group with equal number of final configurations. Thus, a configuration in turn was considered as a target ($Conf_T$) against which the remaining configurations of all the groups ($Conf_i$) were evaluated on fitness (f_i) and on distance (d_i) among their associate genotypes (see Equation 4).

$$
\begin{aligned}
d_i &= H(ind_i, ind_T) \\
f_i = f(Conf_i) = Eval(Conf_i, Conf_T) &= \sqrt{(\Delta\mathscr{A})^2 + (\Delta\mathscr{P})^2 + (\Delta\mathscr{N})^2} \\
\Delta\mathscr{A} &= max\{A_1^T, \ldots, A_m^T\} - max\{A_1^i, \ldots, A_n^i\} \\
\Delta\mathscr{P} &= \sum_{k=1}^{m} P_k^T - \sum_{k=1}^{n} P_k^i \\
\Delta\mathscr{N} &= m - n \qquad (4)
\end{aligned}
$$

Since a configuration comprises a collection of aggregates, a way is needed to perform an evaluation involving all its aggregations collectively. For this reason, considering a target configuration $Conf_T$ and an arbitrary one $Conf_i$ with aggregates $\{A_1^T, \ldots, A_m^T\}$ and $\{A_1^i, \ldots, A_n^i\}$ respectively, $Conf_i$ will be evaluated upon $Conf_T$ in terms of the difference in areas, perimeters and number of achieved aggregations as shown in Equation 4.

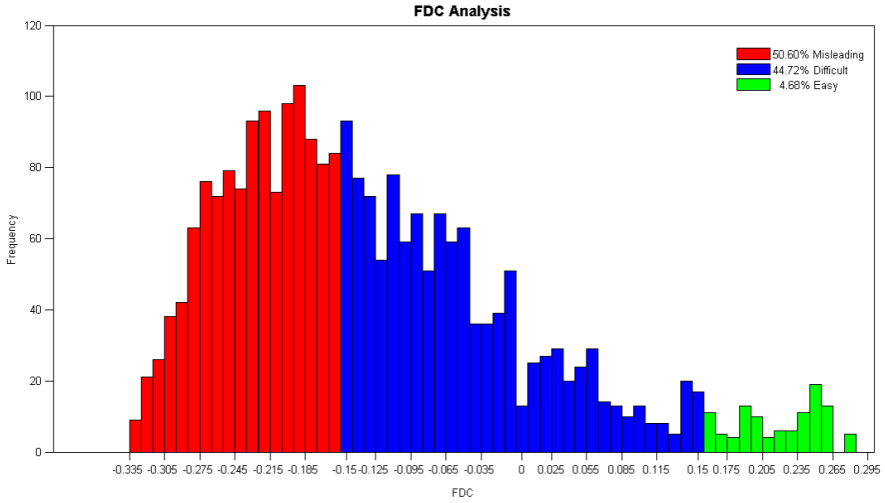

Fig. 5 Proportion of FDC values falling into difficult, misleading and easy to solve categories. From the 2500 analyses performed over 500 individuals, only a 4.68% reveals that a GA may successfully treat the problem.

After performing the calculations, the findings show that the FDC values range from -0.331444 to 0.281457. Since Equation 4 defines a minimisation, 50.60% of the FDC values indicate that using a GA may not be effective, 44.72% that the problem is difficult to solve and a 4.68% that the GA may successfully treat the problem (see Fig. 5).

In particular, visual inspections over scatter plots obtained from the values captured into the smallest percentage depict good correlation on some individuals. A representative sample of these is depicted in Fig. 6 but at `http://www.cs.nott.ac.uk/~gzt/fdcminkowski` we provide the rest of the experiments. For each plot, ind_ij identifies the j-simulation of individual i, where $j \in \{a,b,c,d,e\}$ and $i \in \{1,\ldots,500\}$. Hence, from the sampling of 500 individuals and 2500 simulations subject to FDC analyses, it emerges that employing Minkowski functionals as fitness function offers a relatively satisfactory correlation upon the relationship genotype-fitness for half of the putative samples.

Contrary to the interpretations given by some of the FDC figures seen in this section, the results reported in [11] reveal that using Minkowsi functionals as evaluation method of a GA has positively addressed the self-assembly Wang tiles design problem. Henceforth, we consider that studying and analysing the phenotype-fitness relationship may shed light on the reasons for which such evolutionary approach has been effective.

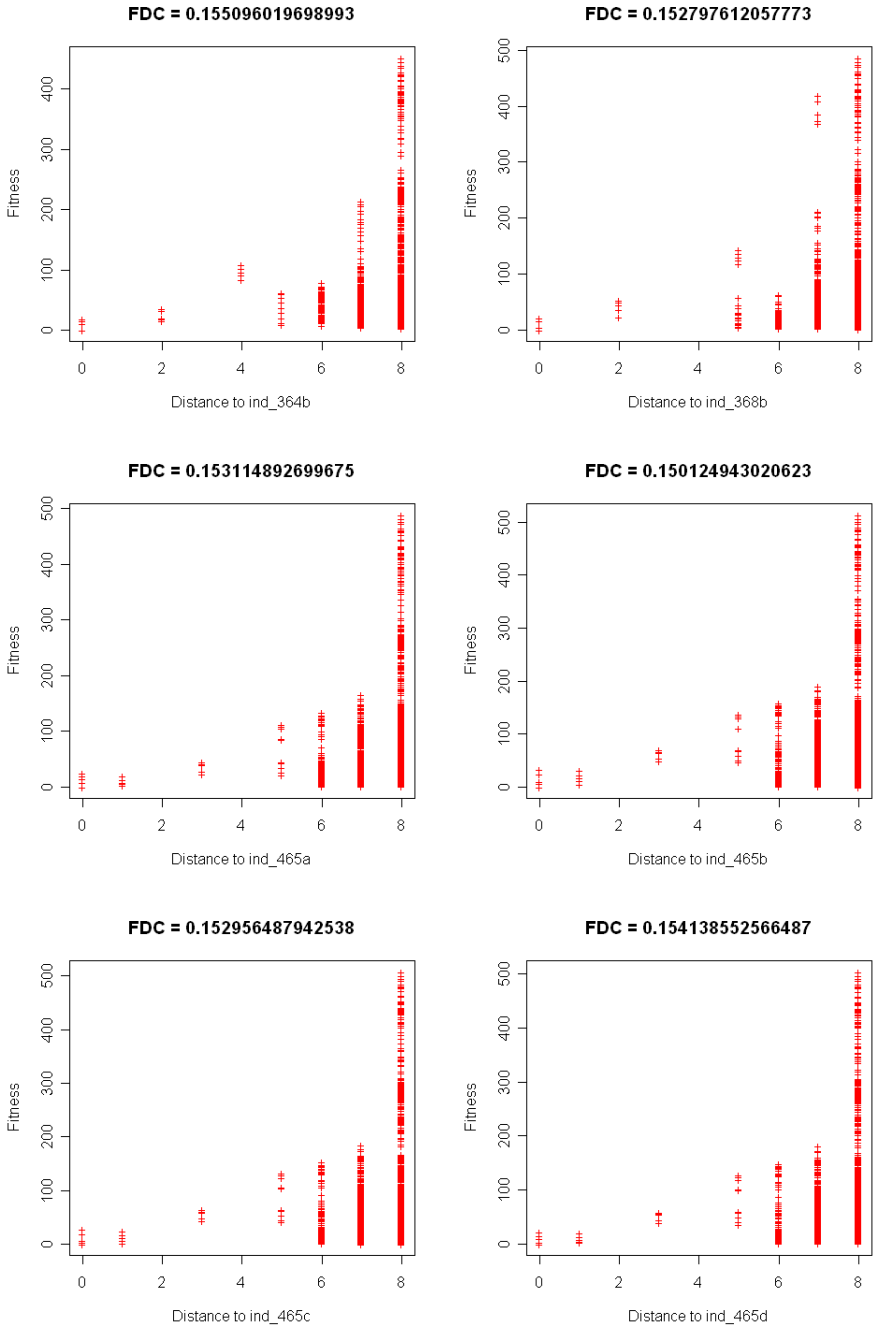

Fig. 6 Graphics of the resultant scatter plots and correlation coefficients for the self-assembly Wang tiles model showing that the Minkowski functionals has a relatively satisfactory correlation with the genotype for some of the self-assembly Wang tile families.

4 Conclusions

This paper has presented an assessment protocol to study the effectiveness of Minkowski functionals as fitness function of our GA employed for the design optimisation of self-assembly Wang tiles.

The results obtained with the morphological image analyses in [11] supports the use of Minkowski functionals as fitness function. However, from the systematically obtained individuals, only 5% of the FDC values have revealed that our GA may successfully solve the problem and 44.72% that the problem could be difficult to solve. Such is the complexity of the genotype-phenotype-fitness mapping, that clearly FDC cannot, alone, be guaranteed to give a completely accurate picture. Indeed, the objective function itself is also only an approximation of two individuals' phenotypic similarity. For these reasons, we conclude that relying on only FDC to validate complex problems would not be adequate. Therefore, we see here the necessity of combining FDC with another method to show with better accuracy whether a given fitness function is a suitable evaluation mechanism for the evolutionary design problem addressed in [11].

As a general conclusion, the application of the methodology shown in Section 3 reveals that employing a fitness function in terms of Minkowski functionals for the evolutionary design optimisation of self-assembly Wang tiles results in a complex mechanism of evaluation where, although its success as phenotype evaluator seems to be appropriate, a more robust analysis is needed for an assessment of how effectively an individual correlates to its genotypic distance to a known optimum. Therefore, the application of such methodology before starting long and expensive evolutionary runs should be considered in any problem where the genotype-phenotype-fitness mapping is complex, stochastic, many-to-many and computationally expensive.

5 Acknowledgements

The research reported here is funded by EPSRC grant EP/H010432/1 Evolutionary Optimisation of Self Assembling Nano-Designs (ExIStENcE).

References

[1] Altenberg L (1997) Fitness Distance Correlation Analysis: An Instructive Counterexample. In: 7th International Conference on Genetic Algorithms, Morgan Kaufmann, San Francisco, CA, USA, pp 57–64

[2] Jones T (1995) Evolutionary algorithms, fitness landscapes and search. PhD thesis, University of New Mexico

[3] Jones T, Forrest S (1995) Fitness Distance Correlation as a Measure of Problem Difficulty for Genetic Algorithms. In: 6th International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 184–192

[4] Koljonen J (2006) On fitness distance distributions and correlations, GA performance, and population size of fitness functions with translated optima. In: Honkela T, Kortela J, Raiko T, Valpola H (eds) 9th Scandinavian Conference on Artificial Intelligence, Finnish Artificial Intelligence Society, Espoo, Finland, pp 68–74

[5] Li L, Siepmann P, Smaldon J, Terrazas G, Krasnogor N (2008) Automated Self-Assembling Programming. In: Krasnogor N, Gustafson S, Pelta D, Verdegay JL (eds) Systems Self-Assembly: Multidisciplinary Snapshots, Elsevier

[6] Michielsen K, Raedt HD (2000) Morphological image analysis. Computer Physics Communications 1:94–103

[7] Michielsen K, Raedt HD (2001) Integral-geometry morphological image analysis. Physics Reports 347:461–538, DOI doi:10.1016/S0370-1573(00)00106-X, URL `http://www.ingentaconnect.com/content/els/03701573/2001/00000347/00000006/art00106`

[8] Quick RJ, Rayward-Smith VJ, Smith GD (1998) Fitness Distance Correlation and Ridge Functions. In: 5th International Conference on Parallel Problem Solving from Nature, Springer-Verlag, London, UK, pp 77–86

[9] Rothemund PWK, Winfree E (2000) The program-size complexity of self-assembled squares (extended abstract). In: 32nd ACM symposium on Theory of computing, ACM, New York, NY, USA, pp 459–468, DOI http://doi.acm.org/10.1145/335305.335358

[10] Terrazas G, Krasnogor N, Kendall G, Gheorghe M (2005) Automated Tile Design for Self-Assembly Conformations. In: IEEE Congress on Evolutionary Computation, IEEE Press, vol 2, pp 1808–1814

[11] Terrazas G, Gheorghe M, Kendall G, Krasnogor N (2007) Evolving Tiles for Automated Self-Assembly Design. In: IEEE Congress on Evolutionary Computation, IEEE Press, pp 2001–2008

[12] Tomassini M, Vanneschi L, Collard P, Clergue M (2005) A Study of Fitness Distance Correlation as a Difficulty Measure in Genetic Programming. Evolutionary Computation 13(2):213–239, DOI http://dx.doi.org/10.1162/1063656054088549

[13] Vanneschi L, Tomassini M (2003) Pros and Cons of Fitness Distance Correlation in Genetic Programming. In: Barry AM (ed) Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference, AAAI, Chigaco, pp 284–287

[14] Vanneschi L, Tomassini M, Collard P, Clergue M (2003) Fitness Distance Correlation in Structural Mutation Genetic Programming. In: Ryan C, Soule T, Keijzer M, Tsang E, Poli R, Costa E (eds) Genetic Programming, Proceedings of EuroGP, Springer-Verlag, Essex, Lecture Notes in Computer Science, vol 2610, pp 455–464