Journal of Biological Physics **27:** 257–283, 2001.

Structural Complexity of Early Embryos: A Study on the Nematode *Caenorhabditis elegans*

F.A. BIGNONE
I.S.T., National Cancer Institute, Lab. Experimental Oncology, Lrgo. Rosanna Benzi, 10, 16123 Genova, Italy, e-mail: abignone@unige.it

Abstract. For analytical studies on the dynamics of gene expression, gene expression control and cellular interactions, the nematode *Caenorhabditis elegans* [*C. elegans*] is at present one of the best suited models [1–4]. In this organism the genetic map and sequence is known [5], moreover the constancy of its lineage tree allows a complete description of cellular clones giving rise to embryos. These characteristics have fostered detailed studies on several aspects of development for this organism. Quantitative studies of cellular movement, through time lapse cinematography of gastrulation, allows the description of cellular migrations giving rise to the final embryonic structure. In perspective, these studies coupled with: genetic analysis, patterns of gene expression obtained through molecular techniques or other methods, open up the possibility of dynamical studies at the organismic scale. This possibility implies, first of all, a study of partitioning of space, and raise several problems in order to define basic conceptual tools to be used in such studies. One of the main problems to handle in this respect is the definition of embryonic structure in a quantitative way. We will show that this aspect is a more general case of distance geometry approaches, as defined in protein folding studies. In this paper we discuss measures of the complexity for embryonal body plans, at the end of grastrulation. These can be applied to studies on the dynamics of gene expression and phylogenetic studies with further experiments or simulations.

Key words: Body plan, *Caenorhabditis elegans*, complexity, development, dynamical systems, gastrulation, genetic networks, protein folding

1. Introduction

Organisms which belong to different Phyla have been known to develop through a strict constant pattern of space partitioning of the fertilized egg with constant cells and nuclei number – cytogeny, eutely[1] –. Species that have been shown to follow this pattern of development, and whose binary trees of replication – partitioning – are known with more or less precision, belong to Phyla as different as Nematoda, Anellida, and Mollusca – Spiralians [6] –.

This phenomenon has been described in details several times during the last century. Early work on the subject has been done in 1890–1901 by A.L. Treadwell [7] using the anellid *Podarke obscura Verril* [*P. obscura*]. More recently detailed descriptions and discussions have been pursued by Bezem and Raven [8] and Raven and Bezem [9] using gastropods. But the experimental system which is

better known is the nematode *C. elegans*, following the seeding work of Sulston and coworkers [2].

These studies have spread the widely held belief of a fundamental difference in the mechanism of development for these species and Phyla, from other Metazoa. Recent evidence, both on *C. elegans* and on other nematodes, seems to dismiss this view [10–15]. This issue raise further problems in the definition of cellular behaviour during early stages of development and, far from dismissing interest in *C. elegans* as a model system, it implies an even increased importance for this model. In fact, if the constant behaviour observed in this species is only apparent, and due to an increase in precision of a mechanism spread to all Metazoa, development of these nematodes should be anyhow due to regulatory mechanisms, as in higher Phyla. But this implies that the well known precision is possible in a total volume of about 10^5 μm^3, and is becoming apparently more loose in those species with larger size and cell number [15]. This makes it an even more wonderful problem to understand for obvious biotechnological and general purposes, and especially for the implications of the role played by chemical noise and chemical instabilities at small scales, in respect of the well known and generally acceptd, strong deterministic behaviour shown by this species. From a more general point of view it must be further considered that the partitioning described above is in general characteristic of a certain organism, also in those cases where a strict correlation with the stereotyped lineage does not exist, but this is not a strict rule. It is anyhow often possible to correlate portions of the volume of the egg into the perspective areas of body plan of embryos.

The time frame of development that we will make reference to in what follows is the period up to the end of *gastrulation*. Gastrulation, in all Metazoa, is the layout of the initial body plan. During this period cells derived from the egg give rise to cell sheets that establish the fundamental shape of the embryo (neuro-ectoderm, mesoderm, endoderm). Later development stem from this initial subdivision usually with further enrichment of the basic structure. From a Biological perspective an understanding of mechanisms involved in this process is important per se but also fundamental for several issues still largely debated. To date only 35 different basic body plans are known in the animal kingdom, and these are almost all known to be derived from the Cambrian radiation (540 million years ago), with no new body types introduced by later evolution [6].

From a biological point of view the important issue is how the genetic background of a certain species, and consequent gene expression, is able to drive this process, and determine the final structure. Moreover it is important from the evolutionary point of view, to understand how body plans can change, giving rise to new species, and further, how different paths of subdivision and rearrangements can give rise to similar final morphologies. This general problem has an overwhelming complexity, involving a wide set of particulars about local cellular interactions. We will not discuss this problem here globally. A general view would make the discussion fall into full fledged fields of taxonomic studies, such as Paleonthology,

Cladistic and Phenetics. Even if some of the aspects discussed here will actually get concepts and draw conclusions pertaining to such fields, we will focus our discussion on those aspects which can at present be rationalized on a quantitative basis, with a further restriction on those organisms where the total volume is constant up to hatching. Finally our discussion will be limited to the spatial arrangement and distribution of cells, as we will show this problem is already difficult enough.

For what concern the theoretical aspects, in recent years the issue of complexity in biological systems, and the way it emerges through evolution, has been debated at length in Molecular Biology, Evolution Theory and Dynamical Systems Theory. In Biology the measures proposed, in order to define complexity of an organism, have been derived from different lines of theoretical and experimental considerations. Among those defined experimentally so far we find: DNA complexity – which has been defined in turn in several different ways as: Shannon's coding complexity, genome size, number of unique coding sequences, C_0t values, etc. –, body complexity – defined as body mass, number of cell types, body height, generation time, etc. – [16–18]. Concerning the emergence of complexity in Biological Systems through evolution, this has also been debated at length in Dynamical Systems Theory, at the molecular level [19, 20], or through the comparison of biological complexity with the richness of behaviour shown in spatially extended systems away from equilibrium [17, 21].

Aside from the case of DNA or proteins, or in the case of spatially extended chemical reactions in which the theoretical counterpart has been used in order to define complexity in several instances, a measure of the intricacies of body plans, and tissues, in animals, which ultimately arise from the patterns of gene expression, has been lacking. In this respect it is also worth mentioning the recent publication of monographs dealing with body plan complexity during evolution [16], or from the point of view of theoretical morphology studies [22], in which a formal definition of biological complexity has not been attempted.

In this paper we discuss general issues raised by development, and in order to be able to give a detailed account of a system, we will limit our discussion to gastrulation. This process is in fact simple enough to be treated fairly, at the same time it is complex enough to overlap all the important aspects that need to be kept into account, in a description at the organismic scale, at least in those Phyla for which cell numbers are low. In the following paragraphs we will discuss similarities between this process and protein folding.

2. Materials and Methods

2.1. EXPERIMENTAL DATA

The examples of normal *C. elegans* embryos used in this paper have already been discussed by Schnabel et al. [10], in particular those considered here are embryos #1, #2, #5. Because of slight variability in cell replication synchrony, cell counts for the three embryos at the end of gastrulation where respectively: 383, 368,

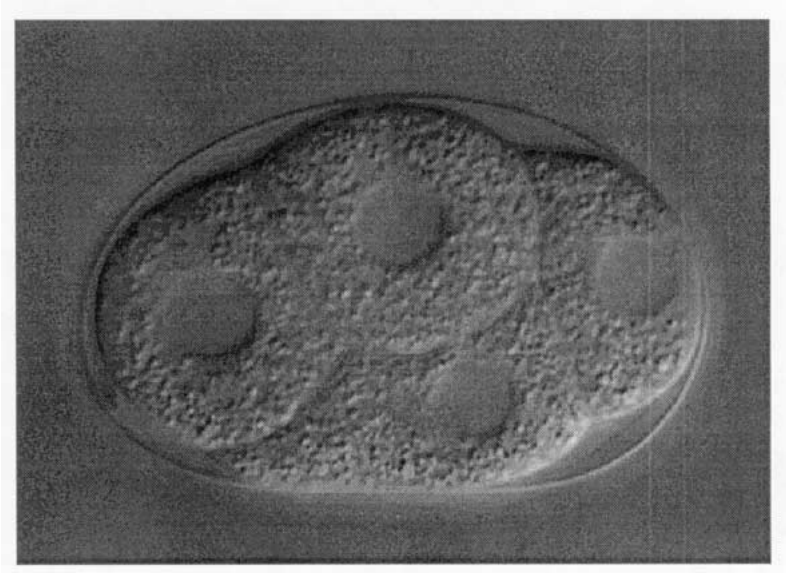

Figure 1. Initial *four cells stage* of a *C. elegans* embryo. From top left going clockwise we find cells named: *ABa, ABp, P*2, *EMS*. Cell *AB* has already gone through an $a - p$ division, in following subdivisions the binary tree of cellular replication will show a similar scheme. The relative constancy of the morphology allows a unique definition of the binary tree and orientation of embryos. The anterior part of the embryo is defined on the side where the *ABa* cell is – left side of the picture shown –, same considerations for the top and bottom sides – top of the picture corresponding to the top of the embryo –.

372 [10]. The experimental data acquisition, through time-lapse cinematography, where carried out with the *C. elegans* wild type strain N2 Bristol cultivated under standard culture conditions [1, 4]. Embryos were grown at a temperature of 25 °C, as previously described [10]. Data were collected by visual analysis of time-lapse cinematography recordings starting from the four cells stage *ABa, ABp, EMS, P*2 –. Microscopy was performed using a microscope equipped with Normarsky optics and a video-camera as described [10]. Data have been collected manually by video manipulation and visual screening. Cellular positions obtained are given as nuclear positions. For this study we make use of three data sets which represent the complete collection of all the cellular position, and cell tagging, of three embryos at the end of gastrulation, as described below.

Data obtained have the folowing format: cell name, time of data collection and three coordinates for the positions (x, y, z). The (x, y) data are taken in VGA video coordinates. They are given as integers and correspond to about $0.1\mu m$ per picture units. The (z) data (depths at the microscope), are given as focal levels (from 1 to 25). These correspond to $1\mu m$ per focus level, with a variability of $\pm 10\%$. The precision for the (x, y) data is thus $10\times$ higher than for the z axis.

For this study we have renormalised the data assuming that the length of the major axis of the embryo has a length of 1.0 – corresponding to about $55\mu m$ –. In

this way variations in the size of different eggs are eliminated, thus measures are given in md, fractions of the major diameter.

2.2. SIMULATIONS AND DATA ANALYSIS

Computations have been done using programs written mostly in C, with few additions of Fortran 77 and Fortran 90. Compilers used have been those for the IBM AIX environment, versions 3.2.5 and 4.2, and the C/C++ GNU compilers for Linux, RedHat distribution 5.2 and 6.1. For some of the plots code has been recompiled and run on Silicon Graphics Octane machines. Displays have been done with public domain utilities such as the HDF data format from N.C.S.A. at the University of Illinois. The colour codings used for cellular 12 *cells stage* are as defined in previous work [10], unless mentioned otherwise.

3. Gastrulation in *C. elegans*

Gastrulation, as defined by embryologists, is both the initial fundamental body plan setting, in terms of spatial rearrangement of the initial egg's volume, and the subsequent formation, in specific cells or cellular clusters depending on the system considered, of an expression pattern in terms of genes activated-inactivated.

As we have discussed in the introduction, two major modes of development have been described so far, a constant mode, in which every portion of the initial space is strictly determined as in *C. elegans* both in spatial position and differentiative pattern; and a variable mode in which volume subdivisions and cellular determinants are less strict – i.e. Amphibians, or the nematode *Enoplus brevis* [13] –. The body plan setting, and the process of gastrulation itself, is the relative movement of one portion of the cells inside the cell mass – engulfing, invagination –, and the final setting of three main sheets. An external sheet of cells making the neuro-ectoderm, an internal sheet the endoderm, and a middle sheet the mesoderm. These three sheets will give rise respectively mostly to: nervous cells-skin, intestine, muscle and choelom (when present). In the case of *C. elegans* the first division of the egg gives rise to two cells, defined by embryologists as: AB and $P1$. Siblings derived from the AB cells – *ABa* and *ABp* in Figure 1 – will give rise to the external structures while the descendants of cell $P1$ – *EMS* and $P2$ cells in Figure 1 – will give rise to the internal constituents. This is achieved through a mixing of movements due to cell replications – with a slide of cells because of the resulting increase in total diameter –, and active cellular movements during intermitotic times. Moreover the binary tree of cell replication in *C. elegans* is desynchronised with the left part – corresponding to cells derived from AB – replicating faster than the right half. At the end of gastrulation the total cells are about 380, making about 256 cells from the AB descendants and 128 cells for the $P1$. This allows a larger surface for the former and facilitate engulfing of the latter.

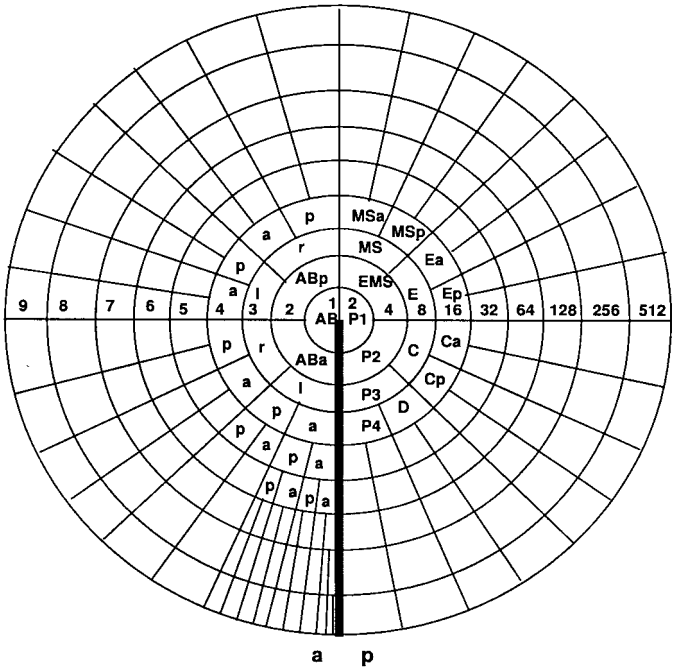

Figure 2. Schematic representation of the binary tree. In the figure the tree lays on branches which show approximately the variation in volume taking place at every cell division. The classical scheme is reported clockwise from center bottom – dark line –. Numbers on the left horizontal line show replications, while those on the right show cell numbers at every level.

The seeding work done by Sulston and coworkers [2] has set the definition of the lineage, following partially historical aspects for name definition of major initial clones, and partially a rationale based on the identification of directions during cellular divisions – anterior a, posterior p, left l, right r, dorsal d, ventral v –. Thus, once the main initial cell divisions are defined – i.e. *four cells stage*, with cells *ABa, ABp, EMS, P*2, see Figure 1 –, the rest of the binary tree is defined with the addition of few further codes for main clones – *C, D, P*3, *P*4 –, and following again a directional scheme at birth.

In this way the name attributed to a cell carries information about its history in terms of location at birth during cellular divisions. For example cell *ABalappaaa*, staring from cell *AB*, has gone through eight divisions, and has been the anterior sister in five divisions, the left in one, and the posterior in two.

In Figure 1 a picture of the *C. elegans* egg is shown at the level of four cells. In this animal typically the four cells stage derives from three cellular divisions the first with a cleavage perpendicular to the major axis of the embryo, and the second and third tilted in respect to it with an $a - p$ bias. As is evident from the picture, cellular divisions at this stage are unequal, giving rise to cells with slight different volumes, and recognizable for their typical morphology. This implies the

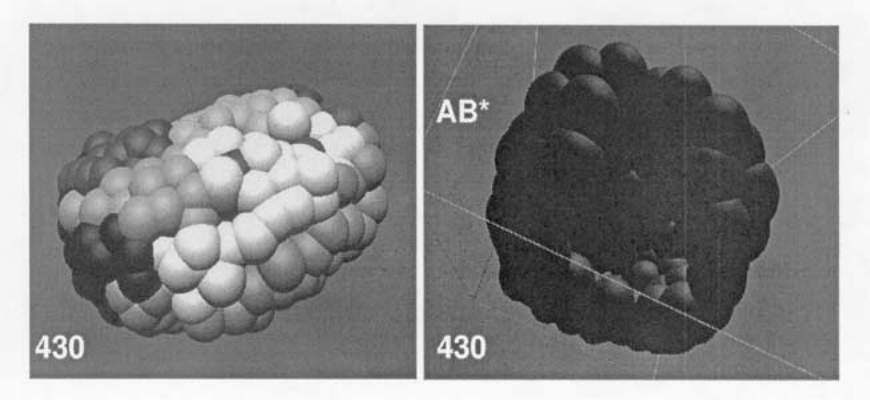

Figure 3. Computer reconstruction of a normal embryo at the end of gastrulation. The colours used for sub clones are as described in Schnabel et al. [10], the clustering of colours implies the absence of mixing between cells which belong to different sub clones, see §5. On the left panel we show the full reconstruction, while on the right we show the empty structure made by the AB* clone, using a unique colour code. This is made by cells derived from the AB cell only, the P1* cells are located inside this structure at the end of the process, and will give rise to the internal structures of the embryo[2].

first important aspect of the model for our discussion: a simple ordering principle for cells in the embryo is possible on a morphological basis. This has allowed in the past the detailed definition of the system.

4. Definition of Cell Threads

The way the cell division pattern of the embryo has been defined allows a study in terms of the structure of the binary tree of replication. In the case of *C. elegans*, and of any other embryo alike, a study of distribution in three dimensions of cells is basically a study on distributions inside a confined volume of tagged spheres. In this case the tagging and final positions are reproducible. At the end of the process, once the cells have been recognized univocally it is thus possible to represent the embryo as a thread, similar to the thread of a protein, linking cells on an ideal line, like a string of pearls. With this reference we can study rearrangements and structure similarly to protein chains.

In Figure 2 we have reported a simplified scheme of space partitioning during cellular divisions. The circle shown in the picture gives an idea in which way cells divide volume wise, and shows the definitions used by embryologists. At the end of the period considered here cells on the right, derived from $P1$, have gone through 7 replications, while cells on the left will have done 8. Distribution of names follows the scheme proposed by Sulston et al. [2] going clockwise from clone *ABala**.

The thread that we define is thus an ideal line connecting all nuclei at their positions, and going as one of the circles in Figure 3 counterclockwise from the most anterior cell defined – i.e. *ABalaaaala* –. This line will correspond, at the end

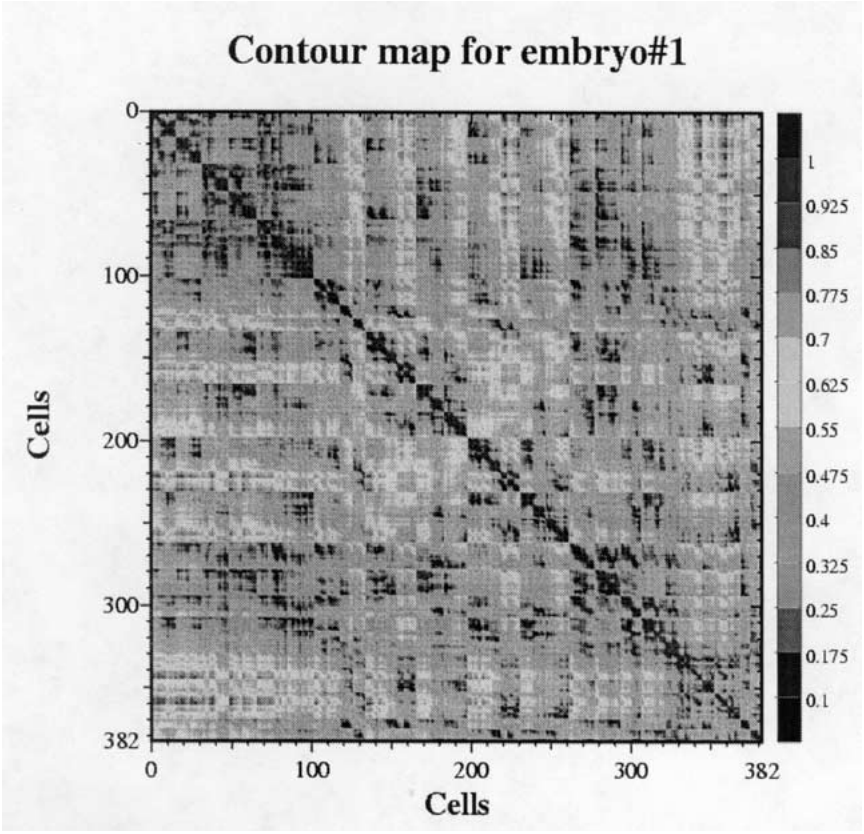

Figure 4. Contour map of embryo #1. All N^2 distances have been calculated based on the measured cellular positions, $N = 383$. The plot has been done colour coding distances after normalization to 1.0 for the maximum distance ($a - p$ maximum length). The scale on the right correspond thus to fractions of the major diameter (md). The dark diagonal imply slight displacement of siblings in respect to the positions of the ancestors.

of gastrulation, to one of the circles in Figure 3, and will be the half circle at the level of 8 divisions for clone AB^*, and 7 divisions for $P1^*$. Twisting, stretching and folding of this ideal thread in space depend on cellular movements during the process, and is a trace of the dynamics.

Cells inside embryos are obviously quite different from atoms or amino acids in a protein, the 'chain' that we consider here does not have a physical base. This implies that the structural situation discussed here is more general in respect of protein folding studies because an important constrain is removed. At the same time the definition that we assume is the most *natural* and follows from the history of cells. Thus positioning ideally cells along the thread sequentially, imply that both the sequence of cellular names and positions follows an arbitrary pattern, but this pattern has good biological motivations and is repeatable. In this way we can achieve a unique description.

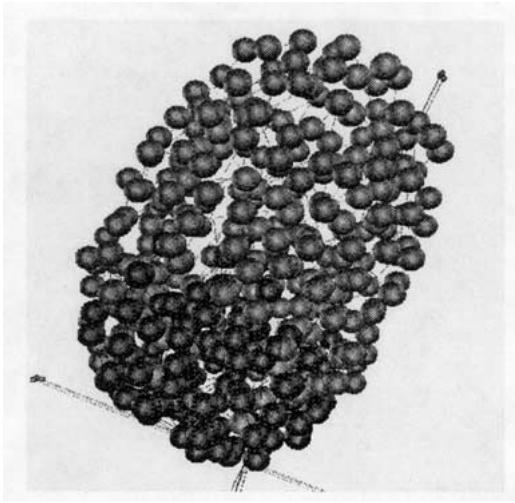

Figure 5. Computer reconstruction of the main subdivision shown in the previous Figure 4, embryo #1. The three major sub clones have been colour coded and plotted using fake spheres centered at the nuclei position. Blue cells represent the anterior *cap* structure, magenta spheres the *middle sleeve*, azure spheres the *interior part*. The all embryo has been tilted counterclockwise for display purposes.

It is worth to emphasize that this experimental setup allows the computer reconstructions of full embryos, for what concern nuclear positions, with a kind of representation similar to those used for Molecular Dynamics [MD] studies. In Figure 3 we show an example of an embryo reconstructed in three dimensions using fake spheres centered at the nuclear positions of its cells.

5. Contour Map and Contact Map

The first possibility in order to study distributions is to build contour maps and contact maps. In the first instance, once all the $d_{i,j}$ distances, between cells i and j have been calculated, one can build a plot as the one shown in Figure 4.

This result shows several important points. The clustering along the diagonal, of values which are among the shortest for the used scale, imply that cognate cells stay in close contact. This is partially obvious, because of the physical nature of cell division. Moreover, there is a clear subdivision of the structure in squared blocks. There are two main interruptions at the level of 130 and 260 cells. These three structures in the embryo make: a *cap*, a *middle sleeve* and an *inner part*, as shown in Figure 5. As we have discussed previously the inner cells are all derived from cell $P1$, the main subdivision between inside-outside is around the level of 260 cells, as discussed above. Further, there is a clear depiction of smaller clusters, up to the level of small groups of cells – groups of 2–4–8 cells –. There is also a tendency to show triangular patterns, this implies a structure for the embryo, seen

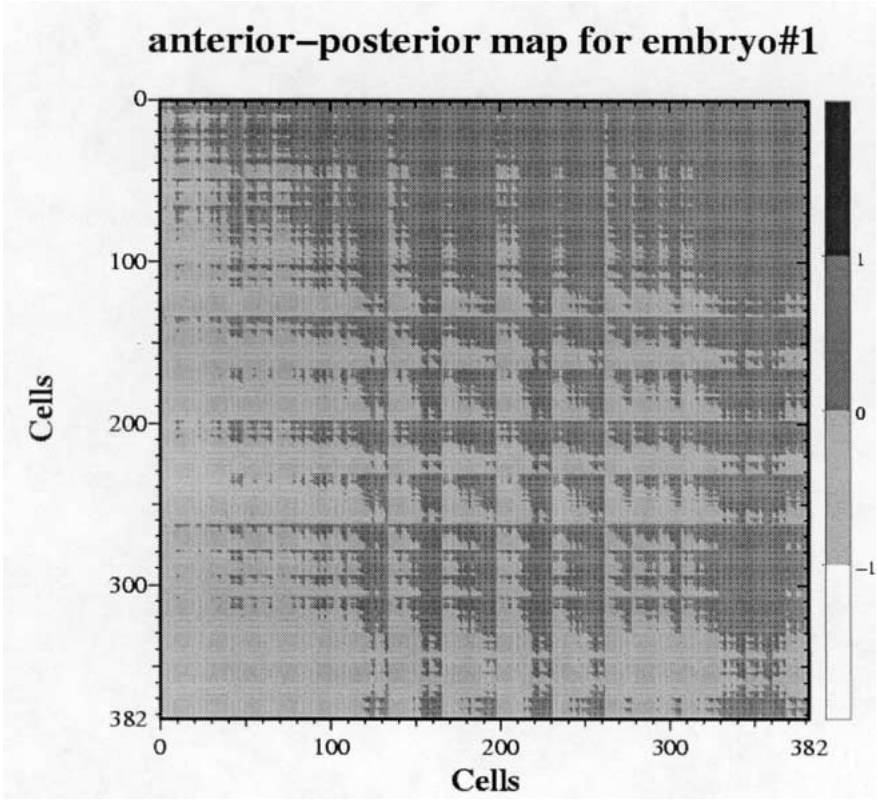

Figure 6. Anterior-Posterior map for embryo #1. This plot has been obtained through a measure of $\Delta x_{i,j}$ given the measured spatial coordinates along the major axis, grey levels in the plot represent respectively positive or negative values for the measure. The triangular shape of the squared regions shows that cells, which are close relatives – sisters and cousins – have a tendency to have an $a - p$ distribution.

as the distribution of cells along a thread, as short sticks. These, made of two-four cells – daughters and cousins –, have a strong anterior-posterior bias.

This fact is shown clearly in Figure 6. In this case we have modified the value of the plotted colour code such that every $d_{i,j}$ is –1 if the calculated difference of positions along the x axis $\Delta x_{i,j} := (x_i - x_j) < 0.0$ and 1 otherwise. The triangular patterns of the plot emphasize the structure discussed, particularly evident for the *middle sleeve* cells.

In order to define *contact maps* we must set some cutoff. For this purpose we rely on the calculation of the pair-distribution [23] for the cells. In our case the definition for $g(r)$ correspond to the form used in MD computer simulations,

$$g(r) = \frac{V}{N^2} \langle \sum_i \sum_{j \neq i} \delta(\mathbf{r} - \mathbf{r}_{i,j}) \rangle \tag{1}$$

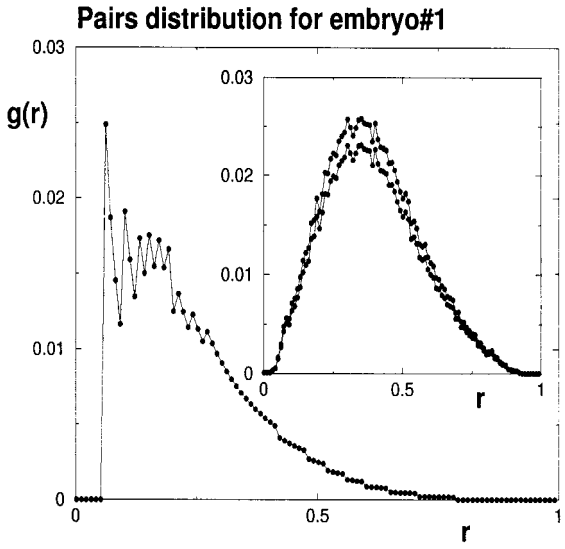

Figure 7. Pairs distribution for cells in embryo #1. In the inlet we show the distribution of the normalized distances **r** as fraction of $1.0md$. The main plot shows the same distribution normalized against the average number of spheres in the same interval in an ideal situation in which all cells are identical and have a density $\rho = 0.85$. See text for further explanations.

Where V is the volume, N is the total number of particles-cells and **r** positions. In our case the measure does not obviously have the statistical character that it acquires in MD, being the result of a single measure of all positions. Thus we have calculated first the bin distribution of the $\binom{N}{2}/2$ distances normalized against N itself. This gives the plot reported in the inlet in Figure 7. After this we have further normalized against an idealized distribution of spheres in space, taking as a reference an ideal gas at a density ρ. This is given by

$$g(r + \frac{1}{2}\delta r) = \frac{n(bin)}{\frac{4\pi\rho}{3.0}[(r + \delta r)^3 - r^3]} \tag{2}$$

where $n(bin)$ is one of the bins of the distribution, and the $\rho = 0.85$ is that for the Lennard-Jones liquid near the triple point. The final result is thus the main plot in Figure 7.

In the case of Biological systems, a general definition of ρ should perhaps imply the study of a particularly regular situation, such as for example taking measures of Drosophila compounds eyes, or some other instance of a particularly regular tissue. But to our knowledge this is the first attempt of a quantitative rationalization of Biological tissue structures with these methods, we had thus to rely on other means.

If one assume a diameter of 1.0 unit for the major diameter of the embryo – $1.0md$ –, and renormalize all positions accordingly, then the distribution of neigh-

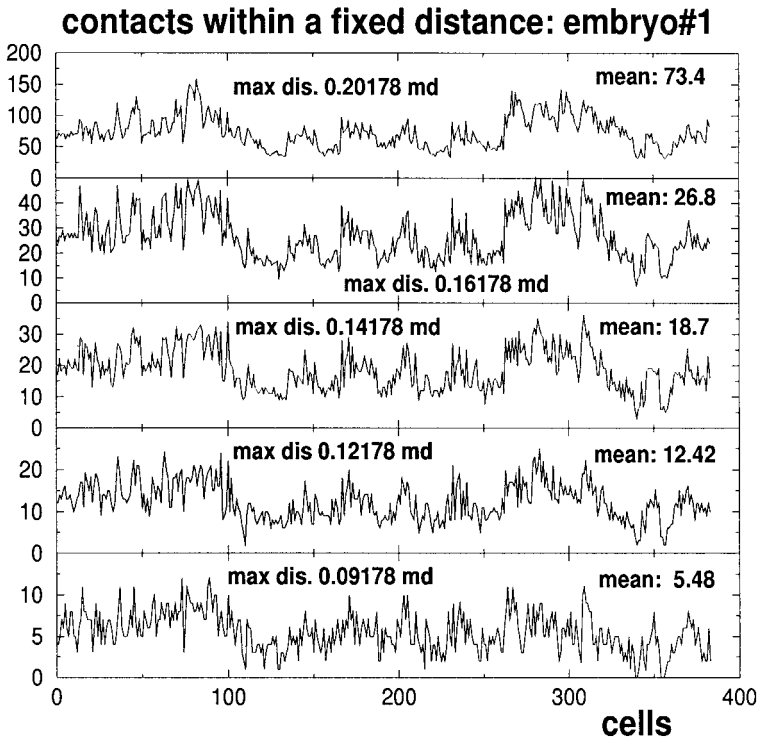

Figure 8. Distributions of the number of neighbours, cell by cell, considering different cutoffs. The cutoff distances chosen for plots are those obtained from the previous figure. The means go from 5.48 to 73.4 as shown. The three qualitative blocks, into which the plot can be subdivided visually, correspond to the areas defined in Figures 4 and 5, because of different cellular mean diameter and general structure.

bours q per cell are those shown in Figure 8 for distances which correspond to peaks in the pair distribution, Figure 7. The number of cells that on average correspond to the first five peaks are respectively: 5.48, 12.42, 18.7, 26.8, 73.4. The first peak, with a small average of neighbours, is due to the boundaries of the system and to the particular positions for smaller cells, sitting as a cell sheet on the outside. It must be noticed that during gastrulation, and in subsequent steps, $g(r)$ is changing during time at every round of replication.

In order to evaluate complexity we have then calculated contact maps using the measures that define different sets of neighbours. The procedure of calculating cells within a certain neighbourhood can be done in different ways. It is possible to calculate a fixed set of neighbours for every cell which are the closest or evaluate the neighbours within a certain cutoff distance. Because of variations in the number of neighbours for a certain cutoff we have considered the mean neighbourhood q for every cell at a certain cutoff. Then we have set at every intersection (i, j) a value of 1 if, for every cell i, cell j is among the closest q cells, and 0 otherwise.

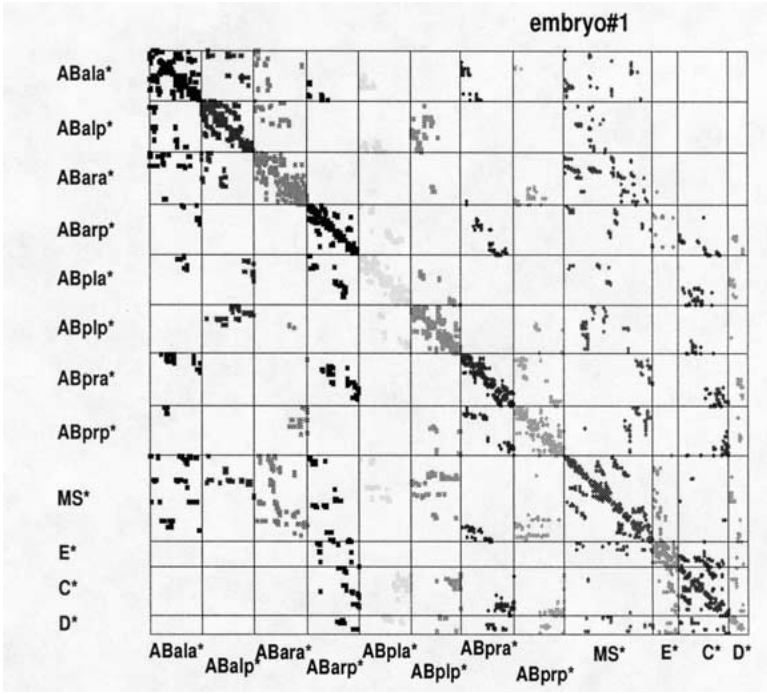

Figure 9. Contact map for embryo #1. The cutoff used has been 0.14178 md, which correspond to a mean of 18.7 contacts per cell – rounded to $q = 19$ –. In the plot the initial 12 major clones are considered, as defined in Schnabel et al. [10]. In order to avoid confusion between cells and clones, the symbol * is used in order to indicate all the cells derived from a certain progenitor. The plot has been done, as in the case of the contour plot shown above, not considering variations in volume among cells, thus plot areas are not proportional to cell volumes but to cell numbers generated by the clones.

With this procedure the resulting contact map can be slightly asymmetric because of local differences in distances. For those measures in which the symmetry of the matrix is important – i.e. the evaluation of spanning trees – the contact map has been instead built by calculating neighbours only for the right upper half of the matrix and then flipping data to the lower left half – i.e. by considering for every i cell those cells in q for which $j > i$ –. In this way a fully symmetric contact matrix is obtained, but in this case the amount of neighbours q per cell is a variable. An example obtained for a cutoff of 0.14178 md with a fixed $q = 19$, is reported in Figure 9. In the plot we report, colour coded, the subdivision of cells into twelve major clones, as defined in Schnabel et al. [10], both for clones definitions and colour codes.

The distribution of contacts is characteristic of the structure of *C. elegans*, and being built on the binary tree of division is dependent on the clonal history of every cell. The structure obtained is thus typical both of the dynamical aspects of gastrulation – cellular divisions and movements –, and of the final positions.

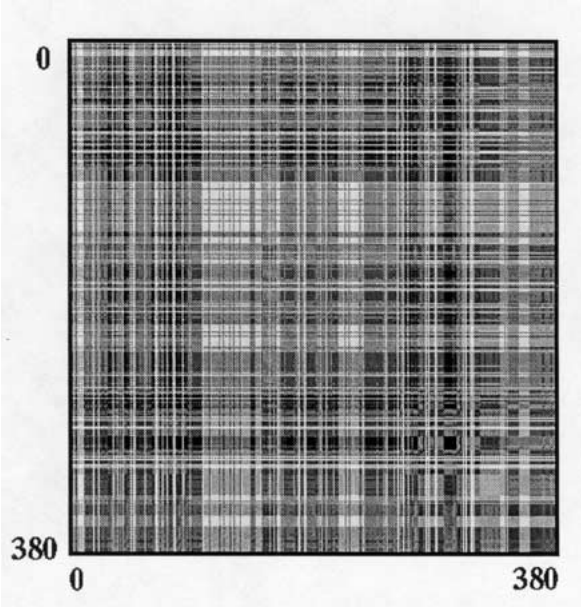

Figure 10. The plot report a simplified expression map for *C. elegans* embryos – edge colouring of the full distance matrix using the known fates –. It must be noticed that, in respect of the binary tree, this plot has been defined as constant [4]. All fates of daughters, of first and second generation, are known and defined. Colour codes are fake codes which correspond to the different possible combinations of descendants fates of a certain cell. The colour attributed at a certain intersection correspond to combinations (products) of such codes. In order to distinguish better different values, a discontinuous colour coding has been adopted.

In the case in which one construct a perfectly symmetric matrix, the picture shown in Figure 9 is the result of an underlying graph Γ made of a vertex set $V\Gamma := \{v_1, v_2, \ldots, v_N\}$, of nodes represented by the position of cells, and an edge set $E\Gamma := \{e_{i,j}, i \neq j\}$ whose elements are the couples defined by the neighbourhood q_i. The plot of Figure 9 becomes in such a case the adjacency matrix $\mathbf{A}\Gamma$ of such a graph.

It is worth noticing that any *row-column switch* operation will leave the underlying contact graph intact aside from name attributions of the nodes-cells. This implies a generalization in respect of protein structures, because the lack of a backbone allows any possible change.

Before discussing complexity we introduce a final twist related to the gene-expression pattern for the cells. In Biological terms, to understand molecular determinants of structure, we must establish possible correlations with patterns of gene expression – i.e. topology of protein signal sources, nuclei –. The construction of an embryo implies the correct expression of a certain genetic pattern in the correct location, in order to give rise to a fully functional structure. This means that the positions we have measured are correlated with a certain set of expressed genes. The level of development at the end of gastrulation is that in which relat-

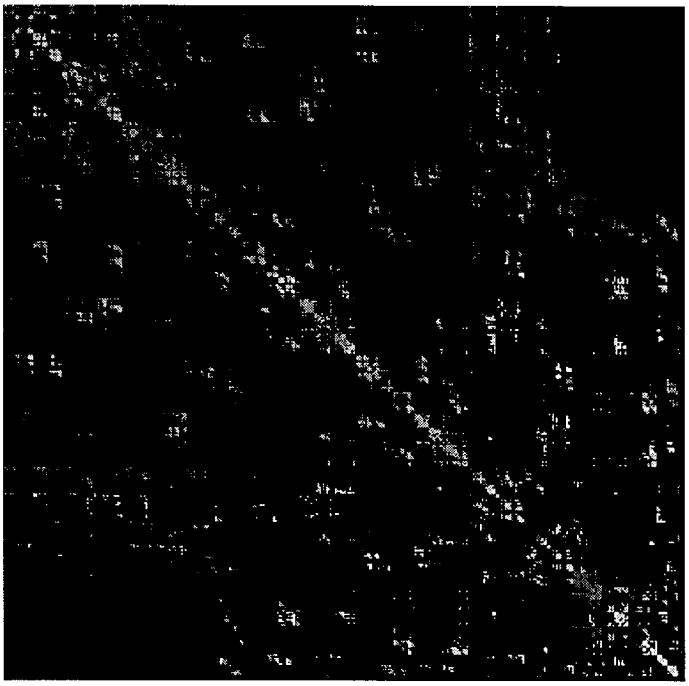

Figure 11. Edge colouring of the adjacency matrix for the contact graph. While the colouring shown above in Figure 9 correspond to an edge colouring derived from a clonal subdivision of the binary tree, in this plot colour codes are derived from a formalization of the differentiative potential. This plot has been obtained as that reported in Figure 10, with a cutoff of 0.14178 md, corresponding to a mean neighbourhood of $q = 19$. There is no simple obvious pattern in the definition of edge colours.

ive positions are basically final, some further movement and replication occurs, shrinking the embryos into a tube, but at the stage of about 380 cells the main body plan, and interrelations among clones in three dimensions, are fixed. In subsequent time steps cells replicate mostly once or twice, with a few exceptions, and give rise to final cell types: neural cells, muscle, cell death .... In order to get a formal code we have attributed to the main differentiation patterns: (Hypodermis, Muscle, Neuron, Pharynx, Intestine, Death, Mitosis) a numerical code. It should be considered that similar patterns could be obtained experimentally with Molecular Biology techniques.

We have then summed codes corresponding to daughter cells of the cell considered, depending on their destiny, and plotted a further product of codes at intersections $df_{i,j}$ of a differentiation-fate map. Results, colour coded, are reported in Figures 10 and 11, identical colours correspond to identical combinations of differentiative destiny of every cells couple $df_{i,j}$ considered. Given a certain gene, or a set of them, correlations with positions can be established for further study. The example shown above is only a pictorial view, based on the known fates of cells,

for discussion purposes related to complexity – see Wood [4] for the complete list of fates used as source here –.

The resulting matrix structure, when considering only a certain cutoff as in Figure 11, is obviously the *edges* colouring for the underlying graph discussed above. In particular in this scheme every edge colour $df_{i,j}$ is a function of vertex colours $v_{c,i}$ and $v_{c,j}$, and in reality cell-cell interactions defined as *inductions* in embryology, are the reciprocal influences between cells which determine a certain vertex colour, which is thus the result of a function of the colours of connected vertices – i.e. local interactions –.

6. Complexity of embryos

As is often the case, also in this instance complexity measures are difficult to define, and depends from which side one wants to tackle the problem. As with most complex systems there are several definitions one can draw concepts and measures from, moreover, the intricacies of early development allows definitions from different points of view.

6.1. RELATIVE CONTACT ORDER

A simple definition, dependent on the *chain type* model discussed above, is the measure of *relative order* [*rCO*]. In protein structure studies *rCO* has been defined as the average sequence distance between all pairs of contacting residues normalized by the total sequence length [24].

$$rCO = \frac{1}{N \cdot Q} \sum^{Q} \Delta S_{i,j} \tag{3}$$

Where N is the total number of residues of a protein – number of cells in our case –, Q is the total number of contacts and $\Delta S_{i,j}$ is the sequence separation in residues – binary replication tree separation, as considered here –, between contacting residues i and j. Results are reported in Figure 12.

The measure is shown increasing the number of contacts by increasing d_{max} – the maximum cutoff distance considered –. Typically for proteins *rCO* starts from values around 0.0747 up to a maximum of about 0.215 [24, 25]. For comparison we have done the same calculation on an embryo in which positions have been shuffled randomly. Results are reported on the top curve of the plot of Figure 12. This can be taken as a maximum for this measure, at about 0.340, because of physical constraints, due to the presence of a backbone, likely no protein can have such value. For the embryos considered this measure move towards this value increasing the number of contacts. For 27 contacts – i.e. Figure 12 – measures are { 0.200; 0.195; 0.195} for the normal embryos of this study, and { 0.325–0.318} for the random one. With this cutoff, methods and measure the complexity of *C. elegans*

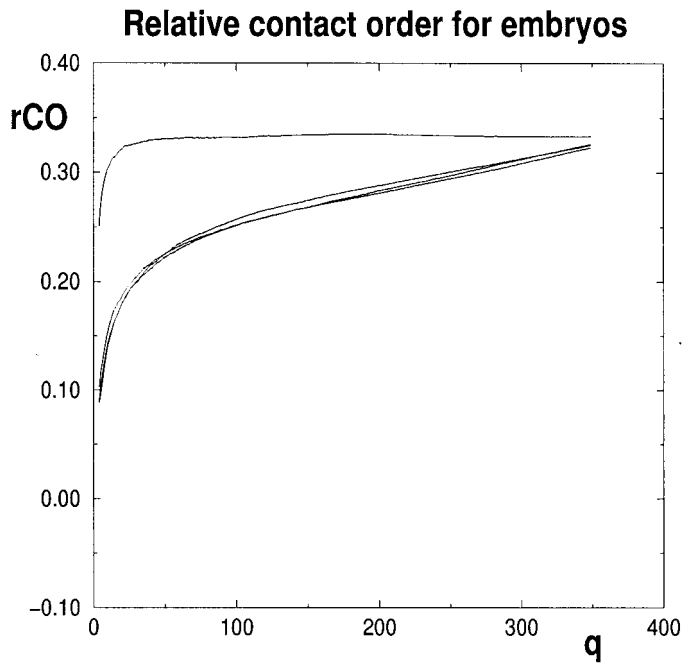

Figure 12. In the figure we report measures of relative contact order *rCO* with a variable number of neighbours, for the three embryos considered. On the abscissa we report the number of neighbours, while on the ordinate the corresponding value for the *rCO* calculated. The picture shows three overlapping plots corresponding to the three embryos considered in this study, the line above is the same measure done on one of the embryos after random shuffling of positions.

embryos is around that of proteins such as *Twitchin* and *muscle acylphosphatase* [25].

The definition of *rCO* can be used also in a slightly different way, but with the same aim. The frequency of sequence separation, of contacts along the thread of cells defined by the binary tree, as defined in the contact map, can be used as a measure of distribution complexity. We rely thus implicitly on the *graph* structure that the contact map imply. If we consider for example the map reported in Figure 9, it is possible to describe embryos complexity on the base of this matrix in two ways. In one case we can consider the distribution of contacts on the map as characteristic of a certain embryo, in the second case we can take the more general view in defining the complexity of the *graph* itself, aside from the tagging of its verticles.

6.2. CONTACTS DISTRIBUTION

In the first instance, similarly to the measure of the *rCO*, we can take as a measure the frequency of distances defined by contacts within a certain cutoff. This is sim-

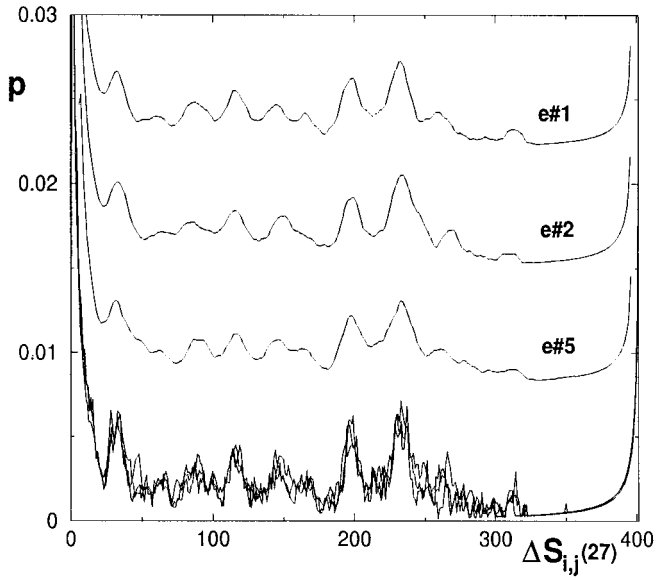

Figure 13. Renormalised curves for probability distributions of neighbours in normal embryos, considering 27 contacts. On the abscissa we report generation distances (binary generation tree distances) – $\Delta S_{i,j}$ –, on the ordinate probabilities. Because of slight differences in cell number – see Materials and Methods –, curves have been rescaled, after computation, for comparison. Running averages for every distribution – 12 points averages – are reported in the plot, labelled according to embryos considered, and shifted vertically.

ilar to the definition of the neighbourhood for Cellular Automata [CA] or Coupled Maps Lattices [CML]. While in CA and CML the neighbourhood is obviously fixed in this case it has a statistical nature. The contact map of a typical neighbourhood as defined in CA and CML, with a regular attribution of vertex tags, would be represented as a set of lines parallel to the main diagonal, aside for few points. Thus, as above, being N the total number of cells and q the number of contacts considered – neighbourhood size –, we have evaluated the frequency for the various cell to cell distances $\Delta S_{i,j}(q)$, normalized against the total contacts

$$p(\Delta S_{i,j}(q)) = \frac{\sum \Delta S_{i,j}(q)}{\sum_{i,j=1}^{N} \Delta S_{i,j}(q)} \tag{4}$$

for every (i, j) couple considered. This count is naturally more frequent for small differences going as

$$p(\Delta S_{i,j}(q)) = \frac{-\Delta S_{i,j} + (N-1)}{N \cdot q} \tag{5}$$

for a random distribution of N elements with q contacts, against which data have been further renormalised. This operation gives an idea of regularities in the un-

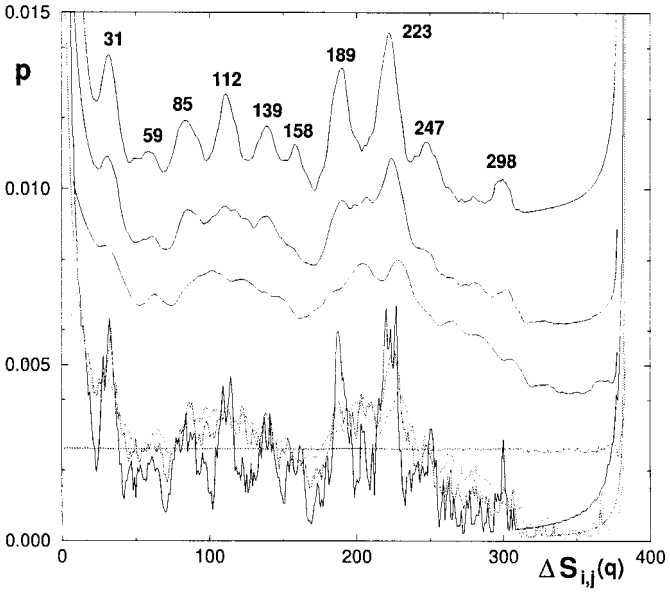

Figure 14. Variations in the profiles, as defined in Figure 13 by changing the number of neighbours considered in the case of embryo #1, neighbours were: 27, 81, 162, 382. Running averages – 12 points – have been reported from top to bottom of the plot as $\Delta S_{i,j}(q)$ against p. Curves have been shifted vertically for clarity, original data are at the bottom of the plot. The horizontal line represent the curve for 382 neighbours.

derlying structure. For the normal embryos considered profiles are reported in Figure 13. These represent the variation in distribution probabilities, in respect of a random shuffling, of finding a certain cell away a distance $\Delta S_{i,j}$, from any cell i along the binary tree.

The plot shows several regularities. The main chance is that of finding a close relative – sisters, cousins – as a neighbour, and this has been already shown in Figures 4 and 9. Moreover, there is a regular increase in probabilities at multiple distances of findings a neighbour there for any cell considered at random. This behaviour results from the fact discussed above – see Figure 6 – of a *stick* structure, made of cell threads which are made of siblings. This reminds of the contact map that one could construct for a regular CML scheme or for a CA, where neighbour sets are defined through a repetitive pattern.

Aside from the first class mostly with neighbours at a distance from $\Delta S_{i,j} = 1$ to $\Delta S_{i,j} = 12$, the other peaks of the distribution – i.e. in the case of embryo #1 – are at values

$$\Delta S_{i,j} = \{31, 59, 85, 112, 139, 158, 189, 223, 247, 298\} \tag{6}$$

with a mean difference of about 30 cells. Thus, aside from the closest neighbours, the structure has six major peaks and four minor ones, making it quasi-two dimensional for contact frequencies. In Figure 14 we have also reported the profile of

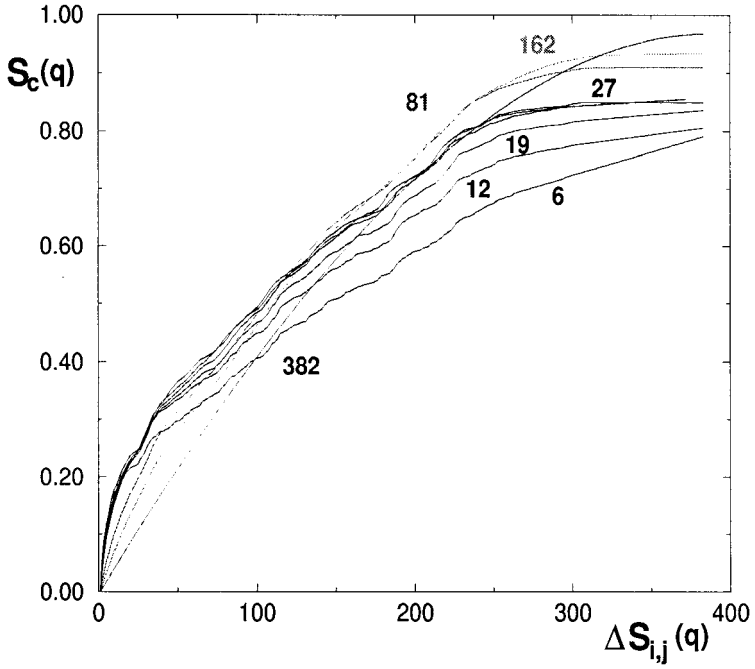

Figure 15. Profiles for $S_c(q)$, as defined in equation 7, with different neighbourhood sizes. Values shown are for embryo #1, and for $q := \{ 1, 12, 19, 27, 81, 162, 382\}$.

frequencies increasing the number of neighbours considered from 27 up to 162. Increasing the cutoff influences smear off as expected.

Following this train of thought one step further it is possible to investigate how changes in position can influence this distribution. This can be done by measuring the distribution itself in terms of complexity. To do so we have evaluated the complexity, for a fixed amount of neighbours q, as

$$S_c(q) = \frac{-\sum p_{\Delta S_{i,j}(q)} ln\, p_{\Delta S_{i,j}(q)}}{\ln(q)} \tag{7}$$

This measure gives results reported in Figure 15 for different q. With this tool it is possible to implement annealing procedures in order to find the simplest possible embryo, i.e. an embryo with an extremely regular structure and a low value for $S_c(q)$. We have done a preliminary investigation of this possibility in two ways. First we have systematically switched positions between two cells i and j, for all possible $\binom{N}{2}/2$ couples. In this case the complexity calculated using formula 7 is mostly increasing after a change, only a small fraction of positions can be exchanged with a gain in regularity. The resulting landscape is very rugged.

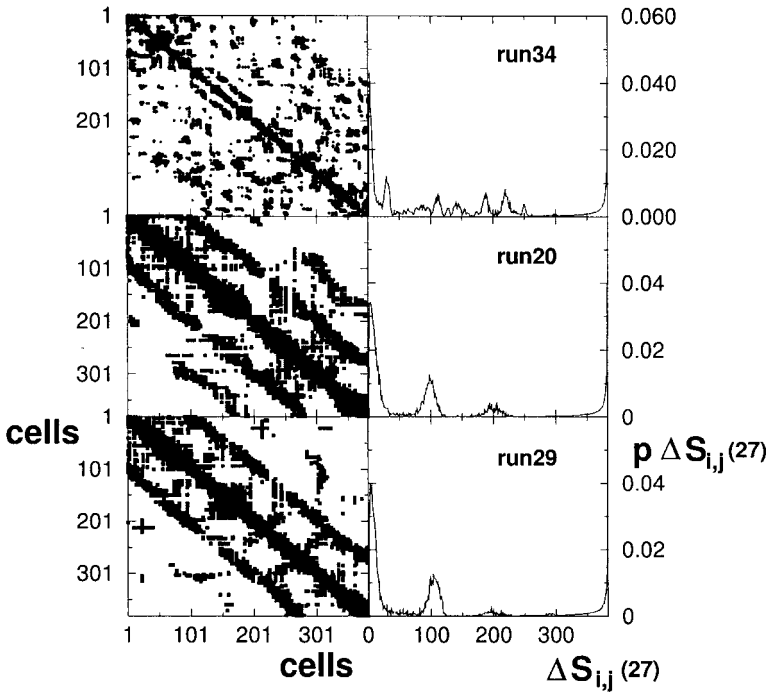

Figure 16. Simple embryos obtained through random searches. In every case shown above two cells have been exchanged in positions at every move. Cells have been chosen in such a way that the second cell of the exchange was always a neighbour, with $q = 27$, of the first choice. A certain move m has been accepted if $S_c(q, m+1) < S_c(q, m)$. Runs have been done respectively for 1.5×10^5 moves for the panel above, and for 3.0×10^5 moves for the second and third panel. The left panels show the final contact maps obtained, while the right panels show the distribution of contacts as in Figure 13.

Then we have performed a random search for regular embryos through local moves only – i.e. changing positions for neighbouring cells – accepting outcomes only for smaller $S_c(q)$ [26]. Some results are reported in Figure 16.

The embryos obtained in this way shows increasing clustering of regular spacing for contacts, with contacts distributions becoming progressively of lower mean dimensionality in respect of real embryos.

6.3. GRAPH COMPLEXITY

A different approach is that of considering the structure of the graph underlying the contact map – adjacency matrix – of embryos. While labelling is formally arbitrary, graph structure is not. Thus we can make use of those properties which are not dependent on graph labels – vertex attributes – and rely only on the graph

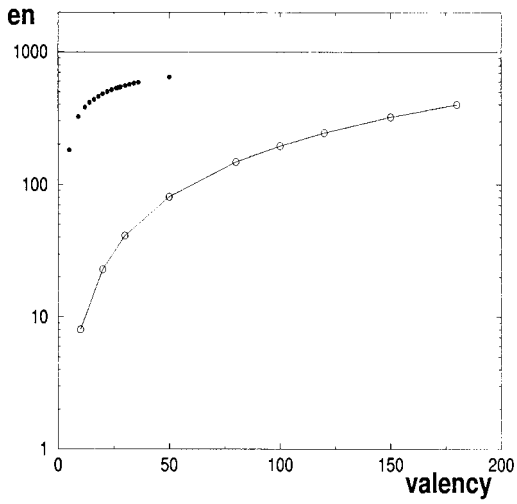

Figure 17. Number of spanning trees for graphs defined with a variable number of neighbours for embryo #1, at the end of gastrulation (383 cells). In the plot we report on abscissa the valency of the graphs while on the ordinate we report the number of trees by plotting only the value of the exponent, on base 10, of resulting complexity measure – number of spanning trees –. The lower curve, labelled with empty circles, represent the complexity for complete graphs – k_n, of valency n – calculated theoretically, (if Γ is complete with n vertices $\mathbf{K}\Gamma = n^{(n-2)}$). The filled dots represent the number of trees for a graph with 383 vertices and a variable number of neighbours – valency –. The horizontal line on the top of the graph shows the number of trees for a k_{383} graph – i.e. a complete graph with 383 vertices: 1.5803E+984 –. To avoid overflow all measures have been done keeping track of exponents on base 10.

structure such as graph spectral properties and graphy complexity. In Graph Theory the complexity $K\Gamma$ of a graph Γ is defined as the number of spanning trees of Γ.

In order to do the calculation, given the adjacency matrix $\mathbf{A}(q)$ – contact map for a certain neighbourhood q –, we have set an incidence matrix $\mathbf{D}\Gamma(q)$ and calculated the augmented adjacency matrix $\mathbf{A}\Gamma_{ug}(q)$ through the product

$$\mathbf{D}\Gamma(q) \times \mathbf{D}\Gamma^T(q) \tag{8}$$

where $\mathbf{D}\Gamma^T(q)$ is the transpose. Then the number of spanning trees of graph $\Gamma(q)$ is the absolute value of the determinant $\mid det \mid$ of any $(N-1)\times(N-1)$ sub matrix of $\mathbf{A}\Gamma_{ug}(q)$ [27]. In this way we have calculated the complexity for a variable amount of neighbours q for embryo #1, results are reported in Figure 17.

7. Discussion

The data and results presented here are an initial formalization of the problem of structure getting in Biological systems, at the organismic level and in an experimental situation. With means currently available it is now possible to build a rationale for the organization of data, in order to give detailed accounts of em-

bryonic structures and the way they emerge during development and evolution, in a quantitative way.

The final goal of this work has been the build up of a conceptual thread, connecting information stored into DNA to the structure and expression patterns in tissues. The problem arises as such because both for descriptive purposes and for manipulation, the level that one must work on is usually that at the DNA sequence level. Moreover, DNA is mostly the base for information storage and retrieval, setting the limits of levels above, and defining a species in genetic terms. But the way in which changes at the DNA level reflect on the organization above is not trivial. Biological systems are in fact a multilevel interaction network:

- The DNA sequence determine both the structure of proteins and the level of their expression through interaction with its local environment. We can wrap in this level both those aspects related to DNA structure, sequence and regulation and RNA-protein structures – i.e. folding –.
- The next level is that of interactions, within a cell and among cells, of gene products and other molecules which give rise to sets of highly interconnected reactions pathways which regulate, and originate from, the level defined above. This level is typically that of spatially extended chemical systems, polimerization and growth phenomena, chemical instabilities, oscillations, biochemical networks. . ..
- Finally the two levels just mentioned, particularly during development but also in adult life, generate and are influenced by the structures formed by cell-cell interactions, cell movement, segregation and selection of shared informations through the cell membrane and membrane proteins, space rearrangements and space partitioning.

The first important point to consider is that all the levels above are interconnected in a two way fashion. Every change at the bottom is reflected at the top, and vice versa. Differently stated the interconnection between parts is tight. Moreover, even knowing the fact that the sequence of DNA is fundamental in setting what is possible for the system and what is not, by determining structure and presence-absence of components, most of the behaviour observed is *emergent behaviour*, i.e. behaviour that results from the dynamical aspects of the system, and by the interactions between constituents given some time and space.

Also for the conceptual tools there is a reflection back and forth of similar aspects. As we have shown, for the definition of embryos, we can rely on theories which have been used for the study of protein folding or for MD studies, and this can, and has been, done also for the intermediate level of biochemical and genetic networks. This is partially obvious because of the nature of theories, but partially it is not so. A global picture of such phenomena imply in fact handling a set of interconnected non-linearities. And this gives rise to theoretical models where the richness of behaviour is overwhelmingly complex [28–31]. With these points in mind, we have followed the path of rationalizing first the complexity present

in a relatively simple experimental situation, before introducing simulations and modeling.

Moreover, problems discussed here have two further facets in methodology, depending on the main immediate goals experimentally. The importance of theoretical aspects is in fact quite different if we want to describe and compare existing experimental systems – different living forms –, or if we want to tackle the problem of prediction of manipulations at the genetic level. In the first case the request from theory is much less daunting, while instead in the second case theory is fundamental. For description it is necessary to go into details first, patterns observed have to deal with particulars related to a certain organism and can be quite specific. For prediction of manipulation the dynamical aspects, emerging behaviour or generic chemical instabilities are instead fundamental, and this is a much more complex game.

For what concern description the Nematode *C. elegans*, as a species, can be defined morphologically and biochemically by maps reported in Figures 4, 9 and 11, even if the fate maps shown are only partial examples. This description can be correlated with the known genomic sequence and biochemical pathways.

Though this description is not complete for several reasons. The variability shown in contact maps and profiles of contact frequencies – Figure 15 –, would require the study of a larger set of embryos. Moreover the final point that we have studied so far is the end point of a story. The next natural steps in the description are going to be a study on the correlations between Figures 4 and 10, and a detailed account of how the structures described here are achieved, i.e. a dynamical description. At the same time tools defined here set the base for such study.

While at the level of sequences – both for DNA, RNA or proteins – the problem has been sized up quite extensively, the reflections that changes in sequences cause on the level of tissue structure and gene expression patterns, when considering these aspects connected, has not been treated in a formalized way until now, also because a lot of biochemistry of these phenomena is still unknown. We have shown here that some formalisms can be introduced and is now possible to correlate in a quantitative way events at quite different dimensional and temporal scales. Thus in perspective it is now possible to establish correlations among the different levels that must be connected for a full study: DNA sequences (including RNA and proteins), Genetics and Metabolic Networks (i.e. interactions among biochemical constituents), Cell-Cell interactions (i.e. cell division, space partitioning, structure building, gene expression patterning).

As far as complexity goes, results obtained have shown that we are dealing with a set of problems which are for certain aspects defined, while for others there is an extension in respect of problems treated extensively so far – i.e. folding –. The rational definition of a species such as *C. elegans* can be done including: genomic information, structural information about biochemical components – DNA, RNA, proteins –, the definition of a binary tree of replication, and a structural graph which include vertex colouring. Theoretically some of this information is redundant, a full

knowledge on genomic information and on dynamical rules should be sufficient to build the all story, but the well-known problems going one step up dimensionally – i.e. folding – tell a different story that goes without comments.

The rationalization presented here can cover all the species in which the binary tree of replication is known with enough precision, together with the differentiative fate, it is thus possible to rationalize comparisons among species both morphologically and in terms of gene expression, knowing enough details. These two aspects given together – definition of maps of Figures 9–11 – open up a pandora's box of combinatorial possibilities theoretically. The absence of a backbone, as in proteins, is in fact increasing possibilities to a full combinatory of the system. At the same time this does not mean functionality for the obtained organism, or other limits such as constraints in cell migrations. We have already pointed out that the known body plans to date are only 35 [6].

Finally, this analysis has been done only on general body plan settings, but it can be extended to any other structure in which enough information is available, so also in subsequent steps of development it is easy to envision the application of this approach to morphological subsets. Thus, while the combinatorics for what concern body plans is apparently strongly restricted, this could be not the case taking a more general point of view. Further, the study of other species and of early mutants, could shred light on the boundaries that we are dealing with and build enough momentum to allow, on the long run, predictive capabilities for genetic manipulations, and experimental testing of different outcomes.

Acknowledgements

Many thanks go to R. Schnabel and H. Schnabel, for stimulating discussions, help in data collection, and for making available for publication the data sets used here. Data presented are part of an ongoing collaborative project still unpublished. The author wish to thanks also R. Livi, A. Politi and A. Torcini, for remarks and discussions, and the I.S.I. Foundation in Torino for hospitality and computing facilities during the initial preparation of this work.

Notes

1. Strictly speaking the terms *cytogeny* and *eutely* refers respectively to the story of the generated cells and to the constancy in nuclei number. While cytogeny and eutely are identical in *C. elegans* during the period of development considered here, this is not always the case during the adult life, being that some cell lineages can give rise to syncytia with slightly different nuclei number. Moreover, these differences can be stronger in other species of nematodes [15]. It must be considered, however, that in several instances in the literature *eutely* is taken as synonymous of cell number, as is often the case.
2. In the paper, we make sometimes use of the symbols AB^* or $P1^*$ to imply all the descendants of that cell. This is done in order to avoid confusion between the cell itself – i.e. $P1$ – and the cell clone to which it gives rise by replication $P1^*$.

References

1. Brenner, S.: The Genetics of *Caenorhabditis elegans*, *Genetics* **77** (1974), 71–94.
2. Sulston, J.E., Schierenberg, E., White, J.G. and Thompson, J.N.: The Embryonic Lineage of the Nematode *Caenorhabditis elegans*, *Dev. Biol.* **100** (1983), 64–119.
3. Riddle, D.L., Blumenthal, T., Meyer, B.J. and Priess, J.R. (eds.): *Caenorhabditis elegans* II, Cold Spring Harbor Laboratory Press, New York, NY, 1997.
4. Wood, W.B. (ed.): *The Nematode Caenorhabditis elegans*, Cold Spring Harbor Laboratory Press, New York, NY, 1988.
5. The *C. elegans* Sequencing Consortium. *Science* **282** (1998), 2012–2018.
6. Raff, R.A.: *The Shape of Life, Development, and the Evolution of Animal Form*, University of Chicago Press, Chicago, Ill, 1996.
7. Treadwell, A.L.: Cytogeny of Podarke Obscura Verril, *J. Morphol.* **17** (1901), 399–460.
8. Bezem, J.J. and Raven, C.P.: Computer Simulation of Early Embryonic Development, *J. Theoret. Biol.* **54** (1975), 47–61.
9. Raven, C.P. and Bezem, J.J.: Analysis of Pattern Formation in Gastropods by Means of Computer Simulation, In: A. Lindenmayer and G. Rozemberg (eds.), *Automata, Language, Development*, North-Holland Publishing Co., Amsterdam, The Netherlands, 1976, pp. 139–145.
10. Schnabel, R, Hutter, H., Moerman, D. and Schnabel, H.: Assessing Normal Embryogenesis in *Caernorhabditis elegans* using a 4D Microscope: Variability of Development and Regional Specification, *Dev. Biol.* **184** (1997), 234–265.
11. Rusin, Yu.-L. and Malakhov, V.V.: Free-Living Marine Nematodes Possess no Eutely, *Dokl. Biol. Sci.* **361** (1998), 331–333.
12. Malakhov, V.V.: Embryological and Histological Peculiarities of the Order Enoplida, A Primitive Group of Nematodes, *Russ. J. Nematol* **6:1** (1998), 41–46.
13. Voronov, D.A. and Panchin, Y.V.: Cell Lineage in Marine Nematode *Enoplus brevis*, *Development* **125** (1998), 143–150.
14. Labouesse, M. and Mango, S.E.: Patterning the *C. elegans* Embryo, *Trends Genet.* **15** (1999), 307–313.
15. Azevedo, R.B.R., Cunha, A., Emmons, S.W. and Leroi, A.M.: The Demise of the Platonic Worm, *Nematology* **2:1** (2000), 71–79.
16. Tyler-Bonner, J.: *The Evolution of Complexity, by Means of Natural Selection*, Princeton University Press, Princeton, NJ, 1988.
17. Kauffman, S.A.: *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford, UK and New York, NY, 1993.
18. Badii, R. and Politi, A.: *Complexity. Hierarchical Structures and Scaling in Physics*, Cambridge University Press, Cambridge, UK, 1997.
19. Eigen, M.: Self-Organization of Matter and the Evolution of Biological Macromolecules, *Naturwissenschaften* **58** (1971), 465–523.
20. Eigen, M., McCaskill, J. and Schuster, P.: The Molecular Quasi-Species, *Adv. Chem. Phys.* **75** (1989), 149–263.
21. Langton, C.G.: Computation at the Edge of Chaos: Phase Transition and Emergent Computations, *Physica D* **42** (1990), 12–37.
22. McGhee, G.R., Jr.: *Theoretical Morphology: The Concept and its Applications*, Columbia University Press, New York, NY, 1998.
23. Allen, M.P. and Tildesley, D.J.: *Computer Simulation of Liquids*, Oxford University Press, New York, NY, 1990.
24. Plaxco, K.W., Simons, K.T. and Baker, D.: Contact Order, Transition State Placement and the Re-Folding of Single Domain Proteins, *J. Mol. Biol.* **277** (1998), 985–994.

25. Grantcharova, V., Alm, E.J., Baker, D. and Horwich, A.L.: Mechanisms of Protein Folding, *Curr. Opin. in Struct. Biol.* **11** (2001), 70–81.
26. Elofsson, A., Le Grand, S.M. and Eisenberg, D.: Local Moves: An Efficient Algorithm for Simulation of Protein Folding, *Proteins* **23** (1995), 73–82.
27. Biggs, N.: *Algebraic Graph Theory*, Cambridge University Press, London, 1974.
28. Bignone, F.A., Livi, R. and Propato, M.: Complex Evolution in Genetic Networks, *Europhys. Lett.* **40:5** (1997), 497–502.
29. Kaneko, K. and Yomo, T.: Isologus Diversification for Robust Development of Cell Society, *J. Theoret. Biol.* **199:3** (1999), 243–256.
30. Bignone, F.A.: Coupled Map Lattices Dynamics on a Variable Space for the Study of Development: A General Discussion on *Caenorhabditis elegans*, *Theoret. Comput. Sci* **217** (1999), 157–172.
31. Hogeweg, P.: Evolving mechanisms of Morphogenesis: On the Interplay Between Differential Adhesion and Cell Differentiation, *J. Theoret. Biol.* **203** (2000), 317–333.