How liquid is biological signalling?

Theoretical Computer Science 410 (2009) 1003–1012

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

How liquid is biological signalling?

Vincent Danos [a,*], Linus J. Schumacher [b]

[a] School of Informatics, University of Edinburgh, United Kingdom
[b] St John's College, University of Cambridge, United Kingdom

ARTICLE INFO

Keywords:
Protein networks
Protein domains
Phase transition

ABSTRACT

This paper proposes an investigation of the global statistics of synthetic protein networks—a step towards a systemic understanding of their design space. We derive a *liquidity index* which describes the onset of the phase transition where an ensemble of agents aggregates into a giant cluster. This index captures the influence of both the domain distribution of agents and the binding strengths of their various domains in the limit of infinite populations. In simple cases it is possible to derive an explicit analytical expression of this index, which allows one to compare with simulations, and get a sense of how it transfers to the concrete finite case.

1. Introduction

Protein networks have been serving the information processing needs of eukaryotes for a billion years, and the modern protein–protein interaction assortment offers a highly combinatorial and flexible toolkit operating on fast time scales [1]. Even so, today's synthetic biology relies mostly on the much slower prokaryotic transcriptional logic. There are various reasons for that. One is that the technology of synthetic protein networks if more efficient is also very demanding. For one thing, eukaryotic cells have complex global structures and constraints one has to deal with (such as shuttling transcription factors to the nucleus [2]). Besides and unlike transcription where a promoter can be considered to regulate the transcription of the genes it is placed upstream of, whatever these genes are, the various domains of a protein interact in elaborate ways that one has only begun to understand [3]; so one also has to deal with sophisticated local constraints on designs. But this is not the only obstacle to the design of large synthetic protein networks. Indeed, even if one were granted a complete operational knowledge of protein domains and their interactive capabilities, one would still lack a way to organize them into an executable description. In fact, in the rare cases when the signal choreography is described in exquisite details with its ballet of receptors, enzymes, adapters, and relays down to the minutest resolution of domains and modifiable residues [4], ordinary modelling methods cannot take on the representational challenge. One could say that protein networks lack a logic, by which we mean an idealised model of protein networks with reasonable descriptive and predictive power. A model that one could regard as a proper design space for synthetic networks is not there yet, or is it?

In an earlier work, we have proposed a stochastic calculus of domain binding and modification – called Kappa – which could be a good start [5–8]. Therein proteins are idealised as sets of sites (aka domains, interfaces, etc.), possibly with internal states to account for post-translational modifications. The dynamics of networks is specified by rules which manipulate domains and bindings explicitly, and as a result highly combinatorial pathways can be formalised, modified and analyzed *in numero* with relative ease—that is to say with an ease that no traditional method affords. It is now well recognised that the decomposition of proteins into domains is a key operator in the structure, plasticity and reprogramming of protein networks [3], and makes the statistics of protein networks far more perspicuous [9] (eg the distinction between single- and

* Corresponding author.
E-mail addresses: vdanos@inf.ed.ac.uk (V. Danos), ljs60@cam.ac.uk (L.J. Schumacher).

0304-3975/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2008.10.037

multi-domain hubs reveals different structural, functional, and evolutionary characteristics). The idealisation we propose is fully congruent with this finer-grained view on protein networks, and shares its concerns: how exclusive or concurrent are bindings between agents, how many different complexes can be formed, how large, for how long, and in how many modified states, and does it matter for the transmission and processing of information? All issues which have clear relevance and impact on the way one wants to design, simulate or otherwise analyze protein networks.

Such key theoretical questions about the global statistics of networks remain mostly unasked, and key physical properties untested—which may be yet another reason, of a more theoretical nature, why synthetic biology today is still mainly transcriptional. The goal of this paper is to progress on these issues. Specifically, we wish to investigate the statistics of networks relating to global connectivity, and set out to derive criticality conditions for the dynamic assembly of protein complexes to diverge into a giant aggregate—or to keep the largest object in the system of negligible relative size. We are also interested in related properties such as the distribution of sizes and complexes. In essence, we are asking: how liquid is biological signalling?

Numerical experiments hint at sharp phase transitions in Kappa rule sets, meaning all or none of the realizations of the underlying stochastic process generate large clusters. Yet given the delicate interplay of protein domain distribution and affinities that controls global connectivity it is unclear how to investigate the issue analytically. Here we are lucky since advances in the study of heterogeneous random graphs due to Söderberg and others [10–13] allow us to define for simple systems a good notion of a liquidity index that furnishes a compact description of the phase transition, and may be a good proxy for the entropy of the limit distribution. In doing so we are following a physical style in exposition, not a mathematically tight one. A further rigorous elaboration of these results might be possible following Ref. [14].

One can easily envision that such a liquidity index as we propose would be useful for the parametrisation and control of large synthetic protein networks—in that most of the time, one would expect such systems to remain in a liquid phase. More generally understanding generic properties of the (idealised) medium delineates the constraints natural and synthetic information processing constructs are working with. That is, to the extent that Kappa-like rule sets convey plausible models of protein networks (natural or synthetic), it seems fundamental to understand the properties of their 'complexomes'. Thus our result should be set in a larger perspective where one wishes to elaborate a statistical vision of information processing in protein networks—seen as a *sui generis* computational medium. The conclusion discusses this matter further.

1.1. Outline

We start the paper with a description of the notion of graphs with sites we are interested in, together with their dynamics as continuous time Markov chains. Since the rule sets that we are able to capture in this first statistical analysis are simple *unconditional* ones, meaning bindings and unbindings happen regardless of the context in which the bond is created or erased, we will not need a full description of our working language. The reader interested in general conditional rules may wish to refer to Refs. [5,6] (or Ref. [8] for a more mathematical exposition).

With this description in place we obtain the equilibrium equations which express the ratio of bindings between our domain types, and this is all one needs to describe the global connectivity structure, since bindings are assumed (approximately) independent. Then we present an elaboration of Söderberg's model and use it to derive the criticality condition; we conclude with an analytic solution of a simple case where one has only two domain types and compare with simulation.

2. The dynamic model

Let us first define the data that we will use to describe our idealised universe of binding agents and derive the equilibrium equations that determine the asymptotic ratio of the various types of bindings.

2.1. Basic data

A *random graph with sites* consists of the following data:
- n the set of nodes
- K the (finite) set of *colours*
- Z the *node* random variable with values in $\mathbb{N}^K$
- for each $a, b \in K$ a *dissociation* constant $\Gamma_{ab} \in [0, \infty]$.

Let us comment on each element of the definition in turn. The first datum n is the number of nodes, and as the goal of this paper is to understand some of the statistical properties of the random graphs defined by the above data (in a way which is explained below) we will consider these properties in the limit of infinite n's. In practice n is not infinite of course, so the result presented here is only of heuristic value for concrete real networks (more about the concrete case later).

The second ingredient is the set K of types of domains an agent can expose. Sometimes domain types will also be called colours, and domains will also be called sites, or stubs. Stubs of the same colour will be indistinguishable from the point of view of the dynamics defined below, however it is important that the reader does not confuse stubs and colours (which are types of stubs). For instance, we will see later that even a universe with two colours and many stubs of either colour can generate interesting statistics.

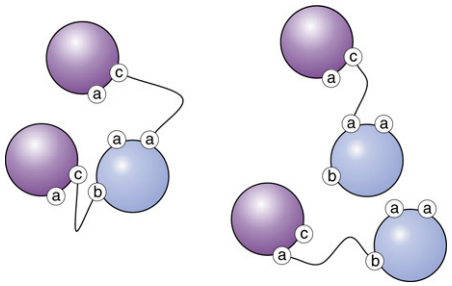

Fig. 1. An example of graph with sites: note that the pairing between stubs/sites may be partial; stub colours are indicated as letters a, b, c.

The next item in the list is a random variable Z which describes how agents are put together as collections of domains; being a random variable it also describes their respective proportions, that is to say $p(Z = m)$ is the probability that a node exposes a collection of domains m. Since there is no mathematical reason to suppose that an agent cannot utilise a same domain type many times, agents are taken to be *multisets* of colours, not mere sets (as would be natural in a concrete biological application). Therefore Z has values in $\mathbb{N}^K$ the set of multisets of colours—seen as maps from colours to their number of occurrences. As above we will use m to denote such multisets of colours, which we sometimes simply call degrees.

Finally the last item is the equilibrium dissociation rate (the ratio of an off-rate and an on-rate, see below) Γ_{ab} which measures the strength of the ab binding. Note that the higher the Γ_{ab} the *weaker* the binding; in particular if $\Gamma_{ab} = \infty$ then one has no binding at all, and obversely if $\Gamma_{ab} = 0$ binding is irreversible. The set of finite Γ's defines the *contact map*—that is the set of domains that may bind to each other. Rates are key to the definition of the limit proportions of edges in our graph as we will see in the next subsection.

We define a (Γ-) *graph with sites* as a finite set of nodes in $\mathbb{N}^K$ together with a partial pairing of their sites which respects Γ, meaning for all pairs of sites x, y of colours a, b, one has $\Gamma_{ab} < \infty$.

Fig. 1 presents a simple example where $n = 7, K = \{a, b, c\}$, and can be drawn according to any node distribution where multisets $a + c$ and $2a + b$ have non-zero probability; one also needs $\Gamma_{ab}, \Gamma_{ac}, \Gamma_{bc} < \infty$.

To simplify our notations we will suppose from now on that no site is self-binding ie $\Gamma_{aa} = \infty$—obviously, this is only a convenience.

For X a set, we write $X^{(2)}$ for the set of unordered pairs of elements of X (equivalently the set of subsets of X with 2 elements), and adhere to the following typographic convention: when an equation is in fact a definition we use the symbol $:=$ or $=:$ depending on which side of the equation is defined.

2.2. Evolution

Given n nodes drawn according to Z, we can now define a continuous stochastic process with values in graphs with sites and of which the dynamics unfolds as follows.

An event can be of two sorts:
- *[binding]* two free sites x, y of respective colours a, b bind each other with a probability proportional to γ^+_{ab};
- *[unbinding]* two sites x, y of respective colours a, b, and already bound together, unbind with a probability proportional to γ^-_{ab}.

The constants $\gamma^+_{ab}, \gamma^-_{ab}$ are respectively called the *ab on-* and *off-rate*.

Write n^f_a for the number of free a sites, and e_{ab} with $a, b \in K^{(2)}$ for the number of edges with ends of colours a, b. The above defines a continuous time Markov chain where the *activity* (or expected frequency) of the system is:

$$\sum_{a,b\in K^{(2)}} \gamma^+_{ab} n^f_a n^f_b + \sum_{a,b\in K^{(2)}} \gamma^-_{ab} e_{ab}.$$

Specifically the activity is the parameter of an exponential distribution which determines the time advance subsequent to an event chosen as explained above. We do not linger too much on the definition of the continuous time Markov chain since this is not useful for the rest (the reader can consult Ref. [15] for definiteness).

One can express the equilibrium dissociation rate Γ_{ab} (see Section 2.1) as $\gamma^-_{ab}/\gamma^+_{ab}$.

As said the above Markov process has in principle the set of Γ-graphs with sites as a state space. But in actuality to describe the system in the limit of large ns, one may forget nodes and only keep track of the number of free sites $(n^f_a; a \in K)$, and edges $(e_{ab}; a, b \in K^{(2)})$. This is because the probability of binding and unbinding events does not depend on the node structure – which is why we have called such systems unconditional earlier – and neither does the activity of the system as one can see in the expression above.

So the two views are statistically equivalent and we choose the latter since it is simpler. To recover the former, ie the limit random graph probability distribution, it is enough to 'glue back' sites according to the degree distribution Z. This will be done implicitly when we define the branching process that explores the size of a connected component to asses criticality (see below).

Writing n_a for the total number of a sites, we get the *invariance* equation (always true if the initial state is in the invariant):

$$n_a = n_a^f + \sum_b e_{ab} \tag{1}$$

where:
- $n_a^f \leq n_a$
- $e_{ab} \leq \min(n_a, n_b)$ is the number of ab bindings.

Note that equivalently e_{ab} is the number of as binding to a b, because we have supposed above that $e_{ab} > 0$ implies $a \neq b$. When stubs of the same colour are allowed to bind one needs to add symmetry factors, a case which we have avoided with our assumption.

2.3. Equilibrium

We can now write the *equilibrium* equation which expresses a stochastic 'equilibrium' (so true for large times and large populations), namely that the dissociation and association activities on ab bindings are the same:

$$\Gamma_{ab} e_{ab} = n_a^f n_b^f. \tag{2}$$

Note that this equation is a property that defines the steady state of the deterministic interpretation of the chain as an ordinary differential equation. We will not do it here, but one may want to relate it formally to the stationary probability distribution of the chain.

We can rewrite the above obtaining a quadratic polynomial system predicting the e_{ab}s as functions of the parameters Γ_{ab}, n_a:

$$\Gamma_{ab} \cdot e_{ab} = \left(n_a - \sum_c e_{ac}\right)\left(n_b - \sum_d e_{db}\right). \tag{3}$$

It is convenient to perform a simple rescaling of the above equations by defining the new parameters $\langle m_a \rangle = n_a/n$, $K_{ab} = \Gamma_{ab}/n$, and the new unknowns $\epsilon_{ab} = e_{ab}/n$.

Note that $\langle m_a \rangle$ is the average number of a sites per node according to Z (this is an approximation which is valid only for large ns obviously), and K_{ab} is the *scale-less* dissociation constant. One can think of the division of Γs by n as a division by a volume term, as is customary when one translates individual-based rates to intensive units (densities or concentrations); likewise one can think of ϵ_{ab} as an edge density.

This gives us an equivalent system over the $K^{(2)}$-indexed variables ϵ_{ab}:

$$K_{ab} \cdot \epsilon_{ab} = \left(\langle m_a \rangle - \sum_c \epsilon_{ac}\right)\left(\langle m_b \rangle - \sum_d \epsilon_{bd}\right) \tag{4}$$

so that a scale-less version of the original data including the colour set K, the node random variable Z, and the K_{ab}s is enough to describe the steady state densities of each edge type.

The former constraint on e_{ab} now translates as $0 \leq \epsilon_{ab} \leq \langle m_a \rangle, \langle m_b \rangle$. This constraint will reappear at the end of the paper, when we solve the equilibrium explicitly in the bicolor case.

We will use the solutions of Eq. (4) to parameterise the static random graph model which we present now. Incidentally, one may wish to prove that this static model defines the limit probability distribution associated to the Markov process just defined. This is an interesting question but we do not need to solve it to proceed—as we shall see.

3. The static model

The data needed to define a coloured degree model (adapted from Ref. [16, p.6]) is the same as in the dynamic model above, except one replaces the dissociation rates Γ_{ab} with a new ingredient Y_a, where $p(Y_a = b)$ is the probability that a stub of colour a binds some stub of colour b.

So one has:
- n the set of nodes
- K the set of colours together with $*$ a special value not in K
- Z the *node* random variable with values in $\mathbb{N}^K$
- for each $a \in K$, Y_a the *edge* random variable with values in $K + \{*\}$.

This new random graph model is *static*, as one no longer describes a stochastic (rewriting) process, but directly defines a probability on a population of graphs with sites. This is the role of the Y_a's. Importantly, we are not supposing that $\sum_{b \in K} p(Y_a = b) = 1$, which amounts to saying that a site may be left free (or unpaired), a fact that we represent by Y_a taking the exceptional value $*$; so $p(Y_a = *)$ may be strictly positive. As dynamic models usually have reversible bindings, one needs to have free sites.

Below we write m_a for the number of stubs of colour a in m (aka the multiplicity of a in m), and $m - b$ for the multiset where there is one less copy of b supposing $m_b > 0$; we also sometimes write simply p_m for $p(Z = m)$.

3.1. The dynamic to static mapping

For the static graph defined by the Y_a's to correspond to the limit behaviour of a dynamic graph as defined in the preceding section (Section 2.3) we need to set the probability that a stub of colour a binds some stub of colour b as:

$$p(Y_a = b) := \epsilon_{ab}/\langle m_a\rangle \tag{5}$$

where the edge density $\epsilon_{ab} := e_{ab}/n$ is given by equilibrium equations (2). The probability that a stub of colour a is connected to one of colour b belonging to a node of degree m is given by $\epsilon_{ab}/\langle m_a\rangle \cdot m_b p_m/\langle m_b\rangle$—since the probability to bind to a b is $\epsilon_{ab}/\langle m_a\rangle$, and the probability that this b belongs to an m is proportional to $m_b p_m$, the probability of m itself and the multiplicity of b in m (in particular if $m_b = 0$ then this is zero as it should).

If we write (note that T_{ab} is symmetric in a and b):

$$T_{ab} := \epsilon_{ab}/\langle m_a\rangle\langle m_b\rangle \tag{6}$$

the above probability is simply $T_{ab} m_b p_m$ an expression which we will use in the next subsection.

3.2. The size generating function

We wish now to evaluate the size of the components in our random graphs. Specifically we are looking for an inductive expression of the random variable S_p^a describing the size of the connected component discovered during an exploration of depth p—which starts exiting from some node by a stub of colour a. That stub could well be free in which case the exploration process stops right away and the discovered size is 0. For a good and customary approximation (see eg Ref. [13]) we will look at this exploration as a branching process which uses alternatively the Y_a's for following an edge to enter a new node, and Z for picking a new exit stub out of the said node. We will study this process using its generating function as well as that of Z:

$$S_p^a(z) := \sum_n p(S_p^a = n) z^n$$
$$Z(x_c; c \in K) := \sum_{m \in \mathbb{N}^K} p(Z = m) \prod_{c \in K} x_c^{m_c}.$$

(To keep the notations light we write the generating function of a random variable as the random variable itself.) Generating functions are a way to display a probability on a countable set that comes particularly handy to study branching processes (see eg Ref. [17]). Note that Z's generating function is a formal power series with a K-indexed set of unknowns, reflecting the fact that the set of values of Z is itself a K-indexed Cartesian product.

Reasoning by cases on the type of node discovered by following the edge (if any) out of a stub of type a, we can express the size generating function S_p^a inductively as:

$$\begin{aligned}
S_p^a(z) - p(Y^a = *) &= z \sum_{n>0} p(S_p^a = n) z^{n-1} \\
&= z \sum_{n>0, m} \sum_{b \in K} T_{ab} m_b p_m \, p\left(\sum_{c \in m-b} S_{p-1}^c = n-1\right) z^{n-1} \\
&= z \sum_{b \in K} T_{ab} \sum_m m_b p_m \left(\sum_{n>0} p\left(\sum_{c \in m-b} S_{p-1}^c = n-1\right) z^{n-1}\right) \\
&= z \sum_{b \in K} T_{ab} \sum_m m_b p_m \prod_{c \in m-b} S_{p-1}^c(z) \\
&= z \sum_{b \in K} T_{ab} \partial_b Z(S_{p-1}^c(z); c \in K).
\end{aligned}$$

On the first line, we have factored out the only case where the size is zero which is obtained when the starting stub of colour a is free with probability:

$$p(Y^a = *) = 1 - \sum_{b \in K} \epsilon_{ab}/\langle m_a\rangle = 1 - \sum_{b \in K} T_{ab}\langle m_b\rangle. \tag{7}$$

On the second line, we use the probability $T_{ab} m_b p_m$ to connect an a to a b belonging to a node of degree m (computed in Section 3.1). The fourth line uses the fact that the generating function of a sum of independent random variables (here the S_{p-1}^c's) is the product of their respective generating functions. The derivation concludes with the introduction of the partial derivative $\partial_b Z$ of the generating function associated to Z (which also has its formal parameters indexed by K).

3.3. Criticality

Define $\mathbf{1} := (1_a; a \in K)$ the node with one stub of each colour.

By induction we can prove that $S_p^a(1) = 1$: firstly, one has $S_0^a(x) = 1$ since in zero jumps one gets zero size with probability 1; secondly, $\partial_b Z(\mathbf{1}) = \langle m_b \rangle$, so by Eq. (7) one obtains $S_p^a(1) = p(Y^a = *) + \sum_{b\in K} T_{ab}\langle m_b\rangle = 1$. This is expected since in general a generating function evaluates to 1 at 1.

Now by taking the limit for large ps in the inductive expression for S_p^a derived above, and fixing z, one obtains a K-indexed system of equations determining the unknowns $(S^a(z); a \in K)$:

$$S^a(z) - p(Y^a = *) = z \sum_{b\in K} T_{ab}\partial_b Z(S^c(z); c \in K). \tag{8}$$

In fact only the particular value $z = 1$ interests us, in which case $\mathbf{1}$ is a solution since as said above $S_p^a(1) = 1$ for all $a \in K$; in other words $\mathbf{1}$ is a fixed point of the function ψ from $\mathbb{R}^K$ to $\mathbb{R}^K$ defined as:

$$\psi_a(x_c; c \in K) := p(Y^a = *) + \sum_{b\in K} T_{ab}\partial_b Z(x_c; c \in K).$$

The spectral radius – by definition the maximum of the absolute values of its eigenvalues – of ψ's Jacobian at $\mathbf{1}$ controls (the convergence of its power sequence, and hence) the stability of $\mathbf{1}$ as a fixed point.

Specifically, if that spectral radius is larger than 1, the fixed point is unstable and there must be a smaller stable fixed point $\eta < 1$. The complement probability $1 - \eta$ is to be understood as the probability that the size of our explored cluster is infinite, i.e. *criticality*. Note that this linear stability analysis will not tell what the relative (infinite) size of the infinite cluster is.

We compute ψ's Jacobian at $\mathbf{1}$:

$$\partial_c \psi_a(\mathbf{x}) = \sum_{b\in K} T_{ab}\partial_c\partial_b Z(\mathbf{x})$$
$$\partial_c \psi_a(\mathbf{1}) = \sum_{b\in K} T_{ab}E_{bc} = (TE)_{ac}$$

where we have written $E_{bc} := \partial_b\partial_c Z(\mathbf{1}) = \langle m_b m_c - \delta_{bc} m_b\rangle$ for the *combinatorial variance* of Z. Hence ψ's Jacobian at $\mathbf{1}$ is simply TE, and $\mathbf{1}$ is stable if the spectral radius of TE, written $\lambda(TE)$ is smaller than 1.

This gives the following criticality condition [12]:

Criticality condition: *An unconditional rule set is critical if* $\lambda(TE) > 1$, *subcritical if* $\lambda(TE) < 1$.

One could regard generally $\lambda(TE)$ as a form of *liquidity index* for a system of interest, in that the closer it gets to 1, the larger the clusters one is likely to observe; when it passes the threshold, one will observe infinite ones in the limit, which in practice means large ones. Of course this is only of heuristic value since the statement above only holds in the limit of infinite systems.

The condition on $\lambda(TE)$ shows clearly the interaction between:
- the node structure, specifically Z's second order moments given by E;
- and the connexion structure given by T.

The dependency in E hints at the fact that node distributions with long tails (typically large hubs) will favour criticality. The contribution of T on the other hand only depends on Z's first order moment and the edge densities ϵ_{ab}. To get a better sense of how this interaction plays out we will explore next the behaviour of our liquidity index in the bicolor case where one can solve the equilibrium equations in closed form. We also finish discussing why our static model must accommodate partial pairings (free sites).

4. Bicolor systems

Let us consider *bicolor* systems with colours a, b where one only allows to pair stubs of opposite colours—meaning $T = \left(\begin{smallmatrix} 0 & t \\ t & 0 \end{smallmatrix}\right)$ for some $t > 0$.

4.1. Numerically

We proceed numerically first and consider the following node distribution:
- $p(Z = 2a) = p(Z = 3b) := \frac{1}{2}$;
- with averages: $\langle m_a\rangle = 1$, $\langle m_b\rangle = \frac{3}{2}$;
- and combinatorial moments: $E_{aa} = 1$, $E_{bb} = 3$ and $E_{ab} = E_{ba} = 0$.

In order to determine the value of t measuring the a, b binding strength from an underlying dynamic random graph model, we first need to fix the corresponding dynamic parameter K_{ab}. Suppose we choose $K_{ab} = \frac{1}{4}$. By Eq. (6) all we need to

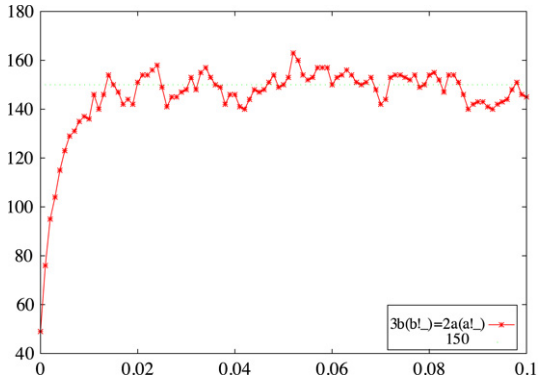

Fig. 2. A run with 100 of each agent types $2a$, $3b$ and $K_{ab} = \frac{1}{4}$. The y axis tracks the number e_{ab} of ab edges over time—for one realization of the continuous time Markov chain (sampled at frequency 10^3 per time unit). The estimated average steady state value 150 – which corresponds to $\epsilon_{ab} = \frac{3}{4}$ – is represented as a dotted line.

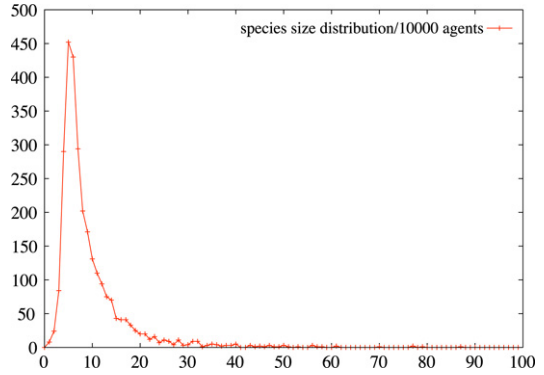

Fig. 3. Histogram of the size distribution of a snapshot at $t = 0.02$ of one realization with 5000 of each agent types $2a$, $3b$ and $K_{ab} = \frac{1}{4}$ (same K_{ab} as in Fig. 2); all clusters are smaller than 1% of the node population.

know is the value of ϵ_{ab} at steady state. This can be obtained either by solving the equilibrium equation (which we do later), or simply by using a simulation (which is an analogous way of solving the equation). We start with the latter. The simulation uses a generic Kappa engine based on an algorithm that extends Gillespie's [18] and is insensitive to the size of the species generated by the system [19].[1]

Looking at the steady state of the simulation for $n = 200$, $\Gamma_{ab} = 50 = nK_{ab}$ we see that on average $e_{ab} = 150$, so $\epsilon_{ab}/\langle m_a \rangle = \frac{3}{4}$ (fraction of bound as), $\epsilon_{ab}/\langle m_b \rangle = \frac{1}{2}$ (fraction of bound bs), and $t = \frac{150}{200} \cdot \frac{2}{3} = \frac{1}{2}$ (Fig. 2). (Of course only the analytic derivation below will show that we read back correctly our parametrisation from the data.)

It is easy to see that $(TE)^2 = \frac{3}{4}I$, so $\lambda = \frac{\sqrt{3}}{2}$ and the system is subcritical. The size distributions observed by simulation are consistent with this; when the population becomes larger ($n = 10\,000$) the size of the largest cluster does not increase in proportion to the population (Fig. 3).

4.2. Aside about partial pairings

This numerical example is a good opportunity to discuss why pairings should be partial, ie why sites in general have to have a non-zero probability to be free. We see that our chosen Z introduces a deficit in as compared to bs, and since T is bicolor this translates in the fact that on average (at steady state) about $\frac{1}{2}$ of the bs are bound whereas $\frac{3}{4}$ of the as are bound. Indeed $2 \times \frac{3}{4} \times 100 = 3 \times \frac{1}{2} \times 100$.

No total pairing has non-zero probability in the sense of Ref. [16] where the probability of a pairing θ is proportional to

$$\prod_{x,y \in \theta} T_{\kappa(x),\kappa(y)}$$

with κ the function mapping each stub to its colour. This is because the diagonal coefficients of T are set to zero, and any total pairing would have to bind two stubs of the same colour. As a matter of fact, it is not possible in general to restrict to

[1] The implementation can be obtained at support@plectix.com.

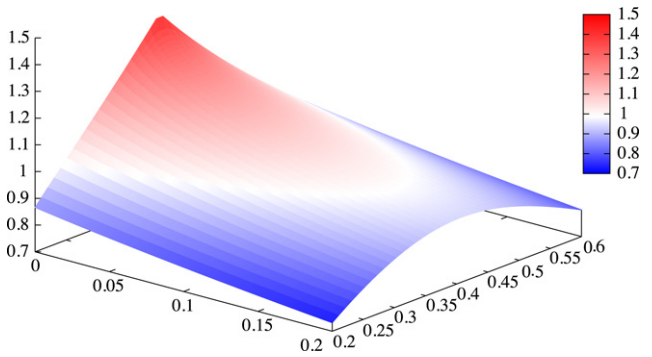

Fig. 4. The front axis is the ratio $0.2 \leq p \leq 0.6$ of ternary agents in the population; the lateral one is the value of the equilibrium dissociation rate K; for $K \leq 0.15$ (strong binding) and $p = \frac{2}{5}$, one has $\lambda > 1$; not so for the values used previously $K = 0.25$, $p = \frac{1}{2}$ (Section 4.1).

total pairings. Fortunately, that poses no problem, as we have seen above, since the criticality analysis carries over as is to the partial case.

On the other hand, partial pairings however do have an impact on T which gets smaller coefficients:

$$\sum_b T_{ab}\langle m_b\rangle = \sum_b \epsilon_{ab}/\langle m_a\rangle =: 1 - f_a \leq 1.$$

With $F := ((1 - f_a); a \in K)$, $T' := F^{-1}T$ satisfies $\sum_b T'_{ab}\langle m_b\rangle = 1$ which expresses the totality of the pairings, so in a way F is a discount that measures the loss of connectivity, hence criticality, incurred by working with partial pairings; and clearly $\lambda(TE) = \lambda(FT'E) \leq \lambda(T'E)$.

4.3. Analytically

Let us now turn to the analytic solution which in the bicolor case is rather easy to obtain and interesting to comment on. The bicolor equilibrium equation is:

$$P(\epsilon) := \epsilon^2 - \epsilon(\langle m_a\rangle + \langle m_b\rangle + K) + \langle m_a\rangle\langle m_b\rangle = 0$$

where we have simply written K for K_{ab}, ϵ for ϵ_{ab}.

Suppose $\langle m_a\rangle \leq \langle m_b\rangle$; since $P(\langle m_a\rangle) = -K\langle m_a\rangle$, $P(\langle m_b\rangle) = -K\langle m_b\rangle$, and $P(\pm\infty) > 0$, the roots $\epsilon^- \leq \epsilon^+$ of the above equation must be such that:

$$[\langle m_a\rangle, \langle m_b\rangle] \subseteq [\epsilon^-, \epsilon^+]$$

which implies in particular that the roots are equal iff $\epsilon^- = \langle m_a\rangle = \langle m_b\rangle = \epsilon^+$ (which also implies $K = 0$).

Because as said earlier $e_{ab}/n \leq \langle m_a\rangle \leq \langle m_b\rangle$, only the smallest root, which we now simply write ϵ, is a meaningful equilibrium:

$$\epsilon := \frac{\langle m_a\rangle + \langle m_b\rangle + K - \sqrt{(\langle m_a\rangle + \langle m_b\rangle + K)^2 - 4\langle m_a\rangle\langle m_b\rangle}}{2}. \tag{9}$$

If in addition we define the following 'noise' term:

$$N := \frac{\langle m_a m_b\rangle + \sqrt{\langle m_a(m_a - 1)\rangle\langle m_b(m_b - 1)\rangle}}{\langle m_a\rangle\langle m_b\rangle} \tag{10}$$

it is easy to see that the system liquidity index is $\lambda = \epsilon N$.

In the simple bicolor case we see how the edge density ϵ and the degree correlation dependent term N separate neatly. Both have a monotone effect on criticality. The higher the density of edges and/or the longer the tail of the degree distribution, the less liquid the system.

One can observe that ϵ only depends on the average coloured degrees $\langle m_a\rangle$, $\langle m_b\rangle$ and the 'pull' K, and that it is clearly a decreasing function of K. When $K = 0$, ie when binding is irreversible, ϵ peaks at the infimum of $\langle m_a\rangle$, stubs of type a are saturated and $\lambda = N\langle m_a\rangle$.[2]

So to drive the system over the transition boundary one can bring K down, but another subtler way to achieve the same is to bring $\langle m_a\rangle$ closer or equal to $\langle m_b\rangle$; which in our numerical example means choosing $p(Z = 3b) = \frac{2}{5}$ to balance on average the number of stubs of each type (more about this in Fig. 4).

[2] When K becomes large $\lambda \sim N\frac{\langle m\rangle^2}{4K}$ where $\langle m\rangle := \langle m_a\rangle + \langle m_b\rangle$.

If we return to the numerical example (Section 4.1), we get:

$$\langle m_a\rangle + \langle m_b\rangle + K = \frac{11}{4},\ \epsilon = \frac{1}{2}\left(\frac{11}{4} - \sqrt{\left(\frac{11}{4}\right)^2 - 4\cdot\frac{3}{2}}\right) = \frac{3}{4}$$

so $e_{ab} = 150$—indeed the value read for the average steady state in the simulation above. And since $N = \frac{2}{3}\sqrt{3}$, $N\epsilon = \frac{1}{2}\sqrt{3} < 1$ and the system is subcritical in accordance with the numerical simulations (Fig. 3) where component sizes stay small relative to the node population.

4.4. Subcritical bicolor systems

A particular, and particularly simple case of bicolor systems is when one has a single agent type bearing one stub of each colour a, b. Connected components are chains, $\langle m_a\rangle = \langle m_b\rangle = 1$, $N = 1$, and $\lambda = \epsilon \leq 1$ which is only critical if $K = 0$. Clearly the probability that a given chain has length k will vary as ϵ^k and decrease rapidly with k. This is in fact a more general phenomenon. If all nodes contain exactly one stub of type a, then the underlying system is subcritical—unless $K = 0$. Indeed, the assumption forces $\langle m_a(m_a - 1)\rangle = 0$ and $N = 1$, so:

$$\lambda = \epsilon = \frac{1 + \langle m_b\rangle + K - \sqrt{(1 + \langle m_b\rangle + K)^2 - 4\langle m_b\rangle}}{2} \leq 1.$$

One sees that the noise term N plays a key role in criticality. Intriguingly, this suggests that large scale polymers made of divalent monomers, because they cannot use too low a K (that would lead to irreversible behaviours), need helper agents which are trivalent or more. Of course this must be taken with a pinch of salt, because biological polymers usually grow in a directed way (and therefore should be idealised by conditional rules which our analytic approach cannot cope with at the moment), and because we are dealing with an idealisation in the first place.

Note that this does not apply to our original $2a, 3b$ example (Section 4.1), and indeed, by choosing carefully the parameters Z, and K, it is possible to obtain critical behaviours; the liquidity index is given by:

$$\lambda(p, K) := \frac{2 + p + K - \sqrt{(2 + p + K)^2 - 24p(1 - p)}}{2\sqrt{3p(1 - p)}}$$

with $p := p(Z = 3b)$ the ratio of $3b$ agents. Plotting λ (Fig. 4) shows where critical behaviour happens.

5. Conclusion

In general a Kappa rule set determines a notion of random graph, namely its stationary probability distribution (under mild assumptions of ergodicity of the underlying Markov chain). Sometimes the state space accessible to a rule set is so large, *and* the dynamics driving the system spread so thinly on the said state space (i.e. the stationary entropy is so large) that it cannot be approximated in any meaningful way by a particular average state. When this is the case—one has to look at a particular state, at least to some extent, as a truly random graph. However if the dissociation rates are large enough the number of significant reachables should be kept low (as a function of the number of agents), and then so will be the limit entropy. In which case, the random graph approach is not useful. It seems key to probe the dividing line in parameter space in between a 'liquid' low-entropy network and a 'solid' high-entropy one—as we have done here in a simple case.

More generally the reason one is interested in such questions is that generic properties of the (idealised) medium make us progress in the understanding of the constraints under which natural and synthetic information processing constructs operate. Besides, and perhaps less ambitiously, the criticality constraint offers a useful sanity check on the dissociation rates of to-be-liquid rule sets. Furthermore and inasmuch as our liquidity index is a good estimate of the dispersion of the set of complexes generated by a simulation, it will give an indication of whether it makes sense to try to enumerate these species, and consequently which simulation technique would fit better the case at hand. Note that not all real networks need be liquid, as there are known examples of solid information processing structures such as *Escherichia Coli's* chemotaxis receptor cluster which seem to undergo a phase transition [20].

To make way in our question and obtain an analysis of criticality, we have used an important restriction to unconditional rule sets. But are unconditional rule sets not too simple to be of any use, the reader will ask. May be they are, but it may also be that the criticality condition will turn out to be a good heuristics for more general rules. One way to understand this would be to extend the treatment to local rule sets in the sense of Ref. [7]. One would also be interested in understanding how conflict, that is the fact that a site type can bind several other types, influences liquidity, or how robust liquidity is (continuity of the condition in the parameters) and which rates will contribute more to its demise (sensitivity). The same question holds for robustness against evolutionary perturbations of a system; that is to say, is it possible to describe plausible transformations for rule sets – perhaps using the notion of rule refinement as in Ref. [8] – and analyze how the criticality condition behaves under such transformations?

Finally it must be said that all the calculations offered within the confines of this note are only about a reasonable yet idealised notion of biological agent, an idealisation which in particular cannot express any serious spatial effects such as molecular crowding, geometric rigidities and steric hindrances. Whether actual synthetic protein networks will mesh well with this idealisation remains to be seen.

Acknowledgements

The first author would like to thank Eric Deeds, Jerome Feret, Jean Krivine, Walter Fontana, and Heinz Köppl for inspirational discussions on the topic of this paper. The second author would like to acknowledge the partial financial support from St John's College (Cambridge).

References

[1] T. Pawson, P. Nash, Assembly of cell regulatory systems through protein interaction domains, Science 300 (5618) (2003) 445–452.
[2] C.M. Ajo-Franklin, D.A. Drubin, J.A. Eskin, E.P.S. Gee, D. Landgraf, I. Phillips, P.A. Silver, Rational design of memory in eukaryotic cells, Genes Dev. 21 (18) (2007) 2271–2276.
[3] R.P. Bhattacharyya, A. Remenyi, B.J. Yeh, W.A. Lim, Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits, Annu. Rev. Biochem. 75 (2006) 655–680.
[4] R.B. Jones, A. Gordus, J.A. Krall, G. MacBeath, A quantitative protein interaction network for the ErbB receptors using protein microarrays, Nature 439 (7073) (2006) 168–174.
[5] V. Danos, J. Feret, W. Fontana, R. Harmer, J. Krivine, Rule-based modelling of cellular signalling, in: L. Caires, V. Vasconcelos (Eds.), Proceedings of the 18th International Conference on Concurrency Theory, CONCUR'07, in: Lecture Notes in Computer Science, vol. 4703, 2007, pp. 17–41.
[6] V. Danos, Agile modelling of cellular signalling, in: Computation in Modern Science and Engineering, vol. 2, Part A 963, 2007, pp. 611–614.
[7] V. Danos, J. Feret, W. Fontana, J. Krivine, Abstract interpretation of cellular signalling networks, in: VMCAI'08, in: LNCS, vol. 4905, Springer, 2008, pp. 83–97.
[8] V. Danos, J. Feret, W. Fontana, R. Harmer, J. Krivine, Rule-based modelling, symmetries, refinements, in: FMSB 2008, in: LNBI, vol. 5054, Springer, 2008, pp. 103–122.
[9] T.O. Yeates, M. Beeby, Proteins in a small world, Science 314 (5807) (2006) 1882–1883.
[10] B. Söderberg, General formalism for inhomogeneous random graphs, Phys. Rev. E 66 (6) (2002) 66121.
[11] B. Söderberg, Properties of random graphs with hidden color, Phys. Rev. E 68 (2) (2003) 26107.
[12] B. Söderberg, Random graphs with hidden color, Phys. Rev. E 68 (1) (2003) 15102.
[13] D. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. USA 99 (9) (2002) 5766–5771.
[14] B. Bollobas, S. Janson, O. Riordan, The phase transition in inhomogeneous random graphs, Random Struct. Algorithms 31 (1) (2007) 3–122.
[15] J. Norris, Markov Chains, Cambridge University Press, 1998.
[16] B. Söderberg, Random graph models with hidden color, Acta Phys. Polon. B 34 (10) (2003) 5085–5102.
[17] D. Williams, Probability with Martingales, CUP, Cambridge, 1991.
[18] D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. 81 (1977) 2340–2361.
[19] V. Danos, J. Feret, W. Fontana, J. Krivine, Scalable simulation of cellular signaling networks, in: Z. Shao (Ed.), Proceedings of APLAS 2007, vol. 4807, 2007, pp. 139–157.
[20] D. Bray, D. Williams, How the melting and "freezing" of protein molecules may be used in cell signaling, ACS Chem. Biol. 3 (2) (2008) 89–91.