DISTRIBUTION OF EDGE COUNTS IN RANDOM MULTIGRAPHS

Abstract

The ChIA-PET technology can identify instances of contacts between different locations of chromosomal DNA sequences as it folds into a three-dimensional structure. Mathematically, this may be modeled as a multigraph where the vertices represent locations in the chromosome and the number of edges represents the number of detected contacts. Such contacts will form randomly as the chromosome folds and unfolds, which can be modeled as random multigraphs. Chromosomal interactions, however, will tend to induce more contacts than otherwise expected, and so the probability distribution for the number of contacts, i.e. the edge counts, provides a means to identify over-represented contacts. We demonstrate that non-central hypergeometric distributions can give accurate estimates of these probabilities, and compute the exact probabilities in order to make the comparison.

1. Introduction

The DNA that makes up the chromosomes folds into a three-dimensional structure. Ongoing studies try to unravel the mechanisms governing this folding and the impact it has on its function. One method for determining the folding structure uses paired-end tag sequencing (ChIA-PET). This uses a protein which can bind to particular DNA segments, e.g. transcription factors, which can be extracted and sequenced using DNA sequencing technologies to determine the exact location in the DNA. The particular version used in ChIA-PET is able to bind to two different DNA segments, which can happen if the two sites are close together in the folded DNA structure, and by determining the location of the two segments, one can identify pairs of chromosomal locations that were in close contact. Since this can be done in a high-throughput manner, one can determine how often such contacts occur. Non-specific contacts can occur through the random motion of the DNA, with a frequency that declines as their distance along the DNA sequence increases. However, specific contacts also arise which are associated with interactions between more distant parts of the DNA.

From a statistical point of view, we may describe this as a set of *contact points*, representing the DNA segments, and a number of contacts between pairs of contact points. A natural mathematical description is as a multigraph, i.e. a graph where pairs of vertices can have multiple edges, in which the vertices are the contact points, and the number of edges between a pair of vertices is the number of detected contacts. Chromosomal interactions should then be expected to give rise to pairs of contact points near the two interacting locations, which can be identified statistically if the number of contacts are unlikely to have arisen through non-specific contacts. This requires a reliable null-model for the frequency of these non-specific contacts.

For contact points p and q, if these only form non-specific contacts, the number of contacts X_{pq} in an experiment should follow a Poisson distributed on the form

$$X_{pq} = X_{qp} \sim \mathrm{Poisson}(\lambda_{pq}), \quad \lambda_{pq} = \alpha_p \alpha_q \eta_{pq} \tag{1}$$

{eqPoisson}

where α_p and α_q are the general binding affinities of the two contact points, and η_{pq} depends on the distance between the contact points measured along the chromosome and represent the frequency with which two contact points with a given distance will come into contact.

We assume that η_{pq} is known. For high-throughput ChIA-PET data, this is a fair assumption since there is sufficient data to allow η_{pq} to be empirically estimated as $\eta_{pq} = h(d_{pq})$ where d_{pq} is the distance between contact points p and q. The binding affinities α_p and α_q, however, are unknown.

If we assume that the number of contacts are given by (1), and the η_{pq} are all known, we can eliminate the influence of the binding affinities α_p by looking at the conditional distribution of (X_{pq}) given the marginal sums $X_p = \sum_q X_{pq}$ which corresponds to the number of contacts for p: i.e.,

$$\mathrm{P}\left[X_{pq} = n_{pq} \,\middle|\, X_i = n_i \text{ for all } i\right]$$

for a specific pair $\{p, q\}$.

The main result is that this conditional distribution can be approximated using Fisher's noncentral hypergeometric distribution

$$\mathrm{hypergeom}(n_{pq}; n, n_p, n_q, \omega) = \frac{C \cdot \omega^{n_{pq}}}{n_{pq}!\,(n_p - n_{pq})!\,(n_q - n_{pq})!\,(n - n_p - n_q + n_{pq})!}$$

where $C = C(n, n_p, n_q, \omega)$ is the normalisation factor which makes this sum to one. The noncentering factor ω is given in (3) which, for cases with large n and many contact points, will tend towards $1/2$.

We provide three arguments in support of this distribution. One is based on conditioning only on X_p, X_q, and the total number of contacts $X = \sum_{\{p,q\}} X_{pq} = \frac{1}{2}\sum_p X_p$, rather than all X_i; this leads to a noncentral hypergeometric model, where we argue that this will give a good approximation to the full conditional model for an appropriate choice of parameters α_i. Another argument uses a Markov chain Monte Carlo process which induces the fully conditional distribution, i.e. where all the marginal sums are conserved, and also leads to an approximation by a noncentral hypergeometric distribution. Finally, in order to properly assess the accuracy of the approximation, we compute the exact distribution for some cases, and compare.

We find that, with the exception of very special cases or very low numbers, the noncentered hypergeometric distrubiton as specified in (2) gives a very accurate approximation of the fully conditional distribution.

2. Statistical model

The problem may be viewed as counting the edges in a random multigraph, i.e. a graph where there may be multiple edges between each pair of nodes, or as a set of random variables $X_{ij} = X_{ji}$ between pairs of contact points forming a random symmetric matrix (X_{ij}). For convenience, we will use some terminology from graph theory, such as node degree to count the number of edges from a node, when that is more convenient. However, most of the time it is more convenient to work with X_{ij}, the number of edges or contacts between i and j, rather than thinking of the individual edges.

Definition 1. Let I be a finite set indexing the node or contact points, and let $I^{\{2\}} = \{\{i,j\} \subset I \mid i \neq j\}$ be the set of pairs. A *random multigraph*, or *contact counts*, is described by a multivariate distribution $(X_{ij})_{\{i,j\}\subset I}$ where $X_{ij} = X_{ji} \in \mathbf{N}_0$. The *marginal counts* are $X_i = \sum_{j\in I} X_{ij}$ where, for convenience, we define $X_{ii} = 0$. The total sum $X = \sum_{\{i,j\}} X_{ij} = \sum_{i\in I} X_i/2$. {*def:count*}

We are interested in the conditional distribution

$$\mathrm{P}\left[(X_{ij}) = (n_{ij}) \bigm| (X_i) = (n_i)\right]$$

where, for convenience, we write (X_{ij}) for the list of values for all unordered pairs $\{i,j\} \subset I$, and (X_i) for the list of marginal sums, etc. Here, $n_{ij} \in \mathbf{N}_0$ are concrete counts, with $n_i = \sum_j n_{ij}$ and $n = \sum_{\{i,j\}\subset I} n_{ij} = \frac{1}{2}\sum_i n_i$ defined as for X_{ij}. In particular, for a given pair $\{p,q\} \subset I$, we want to know the distribution

$$\mathrm{P}\left[X_{pq} = n_{pq} \bigm| (X_i) = (n_i)\right]$$

of X_{pq} given the marginal sums.

There are three different approaches which give rise to the same conditional distribution. The first is a Poisson model where $X_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$ for numbers $\lambda_{ij} = \lambda_{ji} \geq 0$, $\{i,j\} \subset I$. If we condition this on the total count, $X = \sum_{\{i,j\}\subset I} X_{ij} = n$, we get a multinomial distribution $(X_{ij}) \sim \mathrm{Multinom}(n, (p_{ij}))$ where $p_{ij} = \lambda_{ij}/\lambda$, $\lambda = \sum_{\{i,j\}\subset I} \lambda_{ij}$, which indicates the second approach: as a random list of edges, $(e_1, \ldots, e_n)$ drawn from $I^{\{2\}}$, each with probabilities p_{ij}. In the third approach, the conditional distribution arises as the equilibrium distribution of a Markov chain Monte Carlo process on the list $(e_1, \ldots, e_n)$ of edges.

For the Poisson model with parameters $(\lambda_{ij}) = (\alpha_i\alpha_j\eta_{ij})$, denoted P_λ for simplicity, the likelihood function may be split into one factor which depends on the n_{ij} only through (n_i), and another factor which depends on (n_{ij}) but does not contain (α_i):

$$\mathrm{P}_\lambda\left[(X_{ij}) = (n_{ij})\right] = \prod_{\{i,j\}} \frac{\lambda_{ij}^{n_{ij}} e^{-\lambda_{ij}}}{n_{ij}!} = e^{\lambda} \cdot \prod_i \alpha_i^{n_i} \times \prod_{\{i,j\}} \frac{\eta_{ij}^{n_{ij}}}{n_{ij}!}$$

where $\lambda = \sum_{\{i,j\}} \alpha_i\alpha_j\eta_{ij}$. Hence, the conditional distribution $\mathrm{P}_\lambda\left[(X_{ij}) = (n_{ij}) \mid (X_i) = (n_i)\right]$ is invariant under changes in (α_i), and thus depends only on (η_{ij}), and so we may denote this distribution P_η instead of P_λ.

The parameters η_{ij}, λ_{ij}, or p_{ij} have the same effect on the conditional distribution as the non-centering parameter has on the hypergeometric distribution. While in some cases a uniform distribution where all edges are equally likely may be natural, there are important non-uniform cases. E.g., if X_{ij} is the frequency of contacts made between two positions on a string of DNA, a neutral model should not assume all parameters to be equal, but take into consideration that the frequency of random contacts depends on the distance between positions i and j, and incorporate that into the null-model.

3. Distribution conditional on the marginal sums

While the complete conditional distribution can be expressed quite easily as

$$\mathrm{P}_\eta\left[(X_{ij}) = (n_{ij}) \,\middle|\, (X_i) = (n_i)\right] = C_\eta((n_i)) \cdot \prod_{\{i,j\}} \frac{\eta_{ij}^{n_{ij}}}{n_{ij}!}$$

where $C_\eta((n_i))$ is the normalisation constant, the distribution $\mathrm{P}_\eta\left[X_{pq} = n_{pq} \mid (X_i) = (n_i)\right]$ for a given contact $\{p, q\}$ is less easily expressed.

The main result is that

$$\mathrm{P}_\eta\left[X_{pq} = n_{pq} \,\middle|\, (X_i) = (n_i)\right] \approx \text{hypergeom}(n_{pq}; n, n_p, n_q, \hat\omega) \tag{2}$$

{eq:approx}

where

$$\hat\omega = \frac{\hat\lambda_{pq}(n - n_p - n_q + \hat\lambda_{pq})}{(n_p - \hat\lambda_{pq})(n_q - \hat\lambda_{pq})}, \qquad \hat\lambda_{ij} = \hat\alpha_i\hat\alpha_j\eta_{ij} \text{ such that } \hat\lambda_i = \sum_j \hat\lambda_{ij} = n_i. \tag{3}$$

{eq:omegahat}

The solution in $(\hat\alpha_i)$ to $(\hat\lambda_i) = (n_i)$ exists and is unique: see Lemma 1.

3.1. Approximation based on partially conditional distribution

Under the Poisson model, we could relax the conditioning and observe that

$$\mathrm{P}_\lambda\left[X_{pq} = n_{pq} \,\middle|\, X = n, X_p = n_p, X_q = n_q\right] = \text{hypergeom}(n_{pq}; n, n_p, n_q, \omega) \tag{4}$$

{eq:partCond}

has Fisher's noncentral hypergeometric distribution with parameter

$$\omega = \frac{\lambda_{pq}(\lambda - \lambda_p - \lambda_q + \lambda_{pq})}{(\lambda_p - \lambda_{pq})(\lambda_q - \lambda_{pq})}$$

since it corresponds to a 2×2 table of independent Poisson variables: cells contain values X_{pq}, $X_p - X_{pq} = \sum_{i\in I'} X_{pi}$, $X_q - X_{pq} = \sum_{i\in I'} X_{qi}$, and $X - X_p - X_q + X_{pq} = \sum_{\{i,j\}\subset I'} X_{ij}$ where $I' = I \setminus \{p, q\}$. This does not capture the fully conditional model, and depends on the parameters (α_i) used to express $(\lambda_{ij}) = (\eta_{ij}\alpha_i\alpha_j)$. Due to the conditioning in X_p and X_q, however, it does not depend on α_p and α_q.

We can relate (4) to the fully conditional distribution by rewriting it as a sum

$$\sum_{(n'_i)} \mathrm{P}_\eta\left[X_{pq} = n_{pq} \mid (X_i) = (n'_i)\right] \cdot \mathrm{P}_{\hat\lambda}\left[(X_i) = (n'_i) \,\middle|\, X = n, X_p = n_p, X_q = n_q\right]$$

with $(\hat\lambda_{ij}) = (\eta_{ij}\hat\alpha_i\hat\alpha_j)$, and choose $(\hat\alpha_i)$ so as to make $(n'_i) = (n_i)$ the dominant term of the sum. We can do this by setting $(\hat\alpha_i)$ so that $\mathrm{E}_{\hat\lambda}\left[X_i\right] = n_i$: i.e., by solving $n_i = \sum_j \hat\lambda_{ij} = \sum_j \eta_{ij}\hat\alpha_i\hat\alpha_j$, for which the solution in $(\hat\alpha_i)$ exists and is unique (Lemma 1). If the fully conditional distribution $\mathrm{P}_\eta\left[X_{pq} = n_{pq} | (X_i) = (n_i)\right]$ changes only gradually with (n_i), the main contribution to the sum should come from terms with $(n'_i) \approx (n_i)$, for which $\mathrm{P}_\eta\left[X_{pq} = n_{pq} \mid (X_i) = (n'_i)\right] \approx \mathrm{P}_\eta\left[X_{pq} = n_{pq} \mid (X_i) = (n_i)\right]$. And so, this justifies the approximations (2) and (3).

One could ask why conditioning on $X = n$, $X_p = n_p$, $X_q = n_q$ is the natural choice, rather than something less restrictive, or on additional X_i. However, in the Poisson model, we may note that X_{pq} is correlated with X, X_p, and X_q, but is independent from the remaining X_i. Of course, the remaining X_i will be correlated with X_p, X_q, and X, and if some of these dependencies are very strong, that may reduce the accuracy of the approximation.

3.2. Conditional distribution as MCMC equilibrium distribution

Another approach to analysing the conditional distribution comes from obtaining it as the equilibrium distribution under a continuous time Markov chain Monte Carlo process. Central to this is the switching operation in which, for four distinct indices $i, j, k, l \in I$, we can switch between edge combinations $\{i, j\}+$

$\{k,l\}$, $\{i,k\}+\{j,l\}$, and $\{i,l\}+\{j,k\}$ without changing the marginal sums. The random process is such that for any pair of edges, say $\{i,j\}$ and $\{k,l\}$, there is a likelihood $\nu(ij+kl\to ik+jl)\,dt$ for that transition to take place during the next dt of time; if the i,j,k,l are not all distinct, the likelihood is zero. If there are X_{ij} edges of type $\{i,j\}$ and X_{kl} of type $\{k,l\}$, there will be $X_{ij}X_{kl}$ such pairs of edges which may make the transition.

If the transition intensities are the same for all switches, all edge lists $(e_1,\ldots,e_n)$ will be equally likely, which corresponds to all λ_{ij} equal. However, if we pick

$$\nu(ij+kl\to ik+jl)=\frac{1}{\eta_{ij}\eta_{kl}},$$

we get an equilibrium distribution on the list of edges with degree sequence (n_i),

$$\mathrm{P}_\nu\left[(e_1,\ldots,e_n)\right]=C\cdot\prod_{r=1}^{n}\eta_{e_r}=C\cdot\prod_{\{i,j\}}\eta_{ij}^{n_{ij}},$$

with C a normalisation constant. This becomes the equilibrium distribution as it makes all transitions balanced in the sense that they happen equally often in both directions: e.g., if $e_1=\{i,j\}$ and $e_2=\{k,l\}$ where i,j,k,l are all distinct, the transition of e_1 and e_2 to $e_1'=\{i,k\}$ and $e_2'=\{j,l\}$, and the reverse transition, both have frequency

$$\mathrm{P}_\nu\left[(e_1,\ldots,e_n)\right]\cdot\nu(e_1+e_2\to e_1'+e_2')=\prod_{r=3}^{n}\eta_{e_r}=\mathrm{P}_\nu\left[(e_1',e_2',e_3,\ldots,e_n)\right]\cdot\nu(e_1'+e_2'\to e_1+e_2).$$

In addition, it is possible to get from any (n_{ij}) to any other with the same marginal sum using a series of such switches (Lemma 2), which ensures there are not disjoint sets of (n_{ij}) on which the constant C can be set independently. Problems caused by having $\eta_{ij}=0$ can be avoided by setting $\eta_{ij}>0$ and then let $\eta_{ij}\to 0$. If a switch creates an edge $\{i,j\}$ for which $\eta_{ij}=0$, another switch involving this will then take place at once.

Given (n_{ij}), the order of $(e_1,\ldots,e_n)$ can be picked in $n!/\prod_{\{i,j\}}n_{ij}!$ different ways, so that

$$\mathrm{P}_\nu\left[(X_{ij})=(n_{ij})\right]=C\cdot n!\cdot\prod_{\{i,j\}}\frac{\eta_{ij}^{n_{ij}}}{n_{ij}!}\quad\text{when}\quad(X_i)=(n_i)$$

as excpected since it is the multinomial distribution conditional on $(X_i)=(n_i)$.

3.3. Conditional distribution of X_{pq} approximated from MCMC process

The MCMC process produces the conditional multivariate distribution in (X_{ij}) given the marginal sums $(X_i)=(n_i)$. The conditional distribution of X_{pq} alone, however, does not immediately follow as the outcomes of the remaining X_{ij} must be summed over.

We now repeat the MCMC process, but with focus on X_{pq} rather than the whole (X_{ij}). Then, the relevant switch operations are those that change X_{pq}: i.e., those switching between $\{p,q\}+\{i,j\}$ and $\{p,i\}+\{q,j\}$ where $i,j\in I'=I\setminus\{p,q\}$.

The transition intensity $\nu(X_{pq}\to X_{pq}+1)$ is found by summing over all transitions that increase X_{pq} by one:

$$\nu(X_{pq}\to X_{pq}+1)=\sum_{\substack{i,j\in I'\\ i\neq j}}X_{pi}X_{qj}\nu(pi+qj\to pq+ij)=\sum_{\substack{i,j\in I'\\ i\neq j}}\frac{X_{pi}X_{qj}}{\eta_{pi}\eta_{qj}}$$

The transition intensity $\nu(X_{pq}\to X_{pq}-1)$ is similarly found by summing over all transitions that decrease X_{pq} by one:

$$\nu(X_{pq}\to X_{pq}-1)=\sum_{\substack{i,j\in I'\\ i\neq j}}X_{pq}X_{ij}\nu(pq+ij\to pi+qj)=\sum_{\substack{i,j\in I'\\ i\neq j}}\frac{X_{pq}X_{ij}}{\eta_{pq}\eta_{ij}}.$$

The conditional probabilites $p_x=\mathrm{P}_\nu\left[X_{pq}=x\right]$ can now be computed in terms of the transition intensities

$$\nu_x^{\pm}=\mathrm{E}_\nu\left[\nu(X_{pq}\to X_{pq}\pm 1)\;\middle|\;X_{pq}=x\right]$$

using $p_{x-1}\nu^+_{x-1} = p_x\nu^-_x$. We estimate $\nu^\pm_x$ by assuming $|\mathrm{Cov}_\nu[X_{ij}, X_{kl}]| \ll \mathrm{E}_\nu[X_{ij}]\,\mathrm{E}_\nu[X_{kl}]$ for distinct i, j, k, l: in the Poisson model, these are independent, so any correlation is due to the conditioning on $(X_i) = (n_i)$. Next, we use the estimate $\mathrm{E}_\nu[X_{ij}|X_{pq} = x] \approx \mu_{ij}(x)$ where $\mu_{ij} = \beta_i(x)\beta_j(x)\eta_{ij}$ and $(\beta_i(x))$ is selected so as to make $\mu_i(x) = \sum_j \mu_{ij}(x) = n_i$ for $i \in I'$, and $\mu_i(x) - \mu_{pq}(x) = n_i - x$ for $i = p, q$: i.e., similar to (3), but with $X_{pq} = x$ fixed and not included in the estimation of $\beta_i(x)$. This gives us the approximations

$$\nu^+_x = \sum_{\substack{i,j\in I'\\ i\neq j}} \mathrm{E}_\nu\left[\frac{X_{pi}X_{qj}}{\eta_{pi}\eta_{qj}}\middle| X_{pq} = x\right] \approx \sum_{\substack{i,j\in I'\\ i\neq j}} \beta_p(x)\beta_i(x)\beta_q(x)\beta_j(x) = \beta_p(x)\beta_q(x)A(x)$$

and

$$\nu^-_x = \sum_{\substack{i,j\in I'\\ i\neq j}} \mathrm{E}_\nu\left[\frac{X_{pq}X_{ij}}{\eta_{pq}\eta_{ij}}\middle| X_{pq} = x\right] \approx \sum_{\substack{i,j\in I'\\ i\neq j}} \frac{x}{\eta_{pq}}\cdot\beta_i(x)\beta_j(x) = \frac{x}{\eta_{pq}}\cdot A(x)$$

where $A(x) = \sum_{i,j\in I', i\neq j}\beta_i(x)\beta_j(x)$. Thus, using $p_x/p_{x-1} = \nu^+_{x-1}/\nu^-_x$, we get

$$\frac{p_x}{p_0} = \frac{\nu^+_0\cdots\nu^+_{x-1}}{\nu^-_1\cdots\nu^-_x} \approx \frac{\mu_{pq}(0)\cdots\mu_{pq}(x-1)}{x!}\cdot\frac{A(0)}{A(x)} \tag{5}$$

{eq:px-expr}

where $\mu_{pq}(x) = \beta_p(x)\beta_q(x)\eta_{pq}$.

Although it is possible to compute the $\beta_i(x)$ for all i and x, this is cumbersome and impractical. Since $\beta_i(x)$ for $i \in I'$ only enter as part of the bigger sum, we may try to replace it with an average or representative value, $\beta'(x)$. This leads us to replace the equations defining the $\beta_i(x)$ in terms of η_i and $n_{pq} = x$ with the approximations

$$\begin{aligned}\beta_p(x)\beta'(x)(\eta_p - \eta_{pq}) &\approx n_p - x,\\ \beta_q(x)\beta'(x)(\eta_q - \eta_{pq}) &\approx n_q - x,\\ \beta'(x)^2(\lambda - \lambda_q - \lambda_q + \lambda_{pq}) &\approx n - n_p - n_q + x\end{aligned}$$

where the last equation comes from approximating the sum of $\sum_j \beta_i\beta_j\lambda_{ij} = n_i$ over $i \in I'$. From the solution to these approximate equations, we obtain

$$\mu_{pq}(x) \approx \frac{(n_p - x)(n_q - x)}{n - n_p - n_q + x}\cdot\omega, \quad A(x) = \frac{n - n_p - n_q + x}{\lambda - \lambda_p - \lambda_q + \lambda_{pq}}, \quad \omega = \frac{\lambda_{pq}(\lambda - \lambda_p - \lambda_q + \lambda_{pq})}{(\lambda_p - \lambda_{pq})(\lambda_q - \lambda_{pq})}$$

where replacing λ_{ij} with $\hat\lambda_{ij}$ [At some point here I need to switch from λ_{ij} to $\hat\lambda_{ij}$ and this need to be justified more clearly! And earlier?] gives use the ω from (3). Entering these solutions into (5) again gives

$$p_x = \mathrm{P}_\nu[X_pq = x] \approx \text{hypergeom}(x; n, n_p, n_q, \omega).$$

4. Comparisons of approximations to exact conditional distributions

[Add comparisons: figures.]

5. Discussion

[Discuss the two approximations, that derive from the Poisson distribution and the one using MCMC, and explain how they may be considered first and second order approximations: the MCMC approach derives the transition probabilities and only needs the Poisson distribution to estimate the noncentering factor, while the first approach depends more directly on the Poisson assumption.]

Another approximation, presented in [1], reasons that with n edges or contacts, there will be $2n$ contact endpoints. If we consider a random set of contacts from $2n$ endpoints, of which n_p start at node p, to $2n$ enpoints, of which n_q end at node q, without considering that self-contacts within nodes are not permitted and that a contact from i to j is simultaneously a contact from j to i, this results in a hypergeometric distribution Hypergeom$(2n, n_p, n_q)$: i.e., a 2×2 table where the first row sums to n_p, the first column to n_q and the whole table sums to $2n$.

[Compare to cited approximation and explain why they tend to be similar for large numbers.]

Appendix A. Proofs of lemmas

Lemma 1. *Let $\lambda_{ij} = \lambda_{ji} \geq 0$ and $\lambda_{ii} = 0$ for $i, j \in I$, I a finite set, so that $\lambda_i = \sum_j \lambda_{ij} > 0$ for all i. Let $x_{ij} = \alpha_i \alpha_j \lambda_{ij}$ where $\alpha_i > 0$, $x_i = \sum_j x_{ij}$. For any $\mu_i > 0$, $i \in I$, there exists a unique set of values $\alpha_i > 0$ so that $x_i = \mu_i$.* {lemma:aSolve}

Proof. The mapping of $\alpha = (\alpha_1, \ldots, \alpha_n)$ to $x(\alpha) = (x_1, \ldots, x_n)$ is non-singular: i.e. $\det[\partial x_i/\alpha_j] \neq 0$. In particular, for $\alpha_i = e^{a_i}$, we get $\partial x_i/\partial a_j = x_{ij} + \delta_{ij} x_i$, and for any δa_j not all zero we get $\sum_i \delta a_i \cdot \partial x_i/\partial a_j \cdot \delta a_j > 0$.

Given an arbitrary α and the corresponding value x, draw a curve (e.g. a line) $X(t)$ from $X(0) = x$ to $X(1) = \mu = (\mu_1, \ldots, \mu_n)$. We can lift this curve to a curve $a(t)$ with $a(0) = (0, \ldots, 0)$ by using the fact that $\partial[x_i/\partial a_j]$ is invertible whenever $x_i > 0$ for all i. This in turn gives us a corresponding curve $Z(t)$ which maps to $X(t)$, and thus gives $x(A(1) = X(1) = \mu$.

If there are two different values α and α' which both map to the same x, we can draw a curve $A(t)$ from α to α'. This curve maps to a closed curve $X(t)$. Now, we can contract $X(t)$ to a point: i.e. there is a continuous map $X(t; s)$ such that $X(t; 0) = X(t)$ and $X(0; s) = X(1; s) = X(t; 1) = x$. This contraction also lifts to $A(t; s)$, which is impossible since it would require that $A(t; 1)$ be a curve from α to α' for which $x(A(t; 1)) = x$ is constant.

Lemma 2. *Let (n_{ij}) and (n'_{ij}) be two contact counts so that $n_i = n'_i > 0$ for all i. Then, starting with (n'_{ij}), there is a sequence of switching operations that leads to (n_{ij}). A switching operation on (n'_{ij}) consists of taking $\{i, j, k, l\}$ so that $n'_{ij}, n'_{kl} > 0$, reduce n'_{ij} and $n'kl$ by 1, and increase n'_{ik} and n'_{jl} by 1: i.e., the change in contact counts corresponding to replacing edges $\{i, j\} + \{k, l\}$ by $\{i, k\} + \{j, l\}$.* {lemma:switchConn}

Proof. Assume that the contact points are enumerated $1, \ldots, m$. Order the pairs $\{i, j\}$ so that for $i < j$, $k < l$ we say that $\{i, j\} < \{k, l\}$ if $i < k$ or if $i = k$ and $j < l$. Pick the minimal pair $\{i, j\}$ so that $n_{ij} \neq n'_{ij}$.

If $n'_{ij} > n_{ij}$, there must be a k so that $n'_{ik} < n_{ik}$, and this must satisfy $k > j$ since $n'_{ir} = n_{ir}$ for $r < j$. Similarly, there must be an l so that $n'_{kl} > n_{kl}$, which must satisfy $l > i$. Now, the switch $\{i, j\} + \{k, l\}$ to $\{i, k\} + \{j, l\}$ will reduce n'_{ij} by one. We may then go back and repeat the step until we get $n'_{ij} = n_{ij}$.

If $n'_{ij} < n_{ij}$, we may similarly find $k > j$ so that $n'_{ik} > n_{ik}$. There must be an l so that $n'_{jl} > n_{jl}$, which requires $l > i$. Now, switching $\{i, k\} + \{j, l\}$ to $\{i, j\} + \{k, l\}$ will increase n'_{ij} by one. This step may then be repeated until we get $n'_{ij} = n_{ij}$.

After obtaining $n'_{ij} = n_{ij}$ through a sequence of switches, we return to the start to find the next minimal pair $\{i, j\}$ for which $n'_{ij} \neq n_{ij}$. Eventually, $n'_{ij} = n_{ij}$ for all $\{i, j\}$.

Appendix B. Computation of the exact distribution

The exact distribution can be computed from the generating function for the Poisson model. Since, for each $\{i, j\} \subset I$, we have $\mathrm{E}\left[s^{X_{ij}}\right] = \exp(\lambda_{ij}(s-1))$, and the X_{ij} are independent, the combined probability generating function for counting X_{pq} and the marginal sums X_i can be expressed as

$$G_{pq}(u, (t_i)) = \mathrm{E}\left[u^{X_{pq}} \cdot \prod_i t_i^{X_i}\right] = \mathrm{E}\left[\prod_{\{i,j\} \neq \{p,q\}} (t_i t_j)^{X_{ij}} \cdot (t_p t_q u)^{X_{pq}}\right]$$

$$= \exp\left(\sum_{\{i,j\} \neq \{p,q\}} \lambda_{ij} t_i t_j + \lambda_{pq} t_p t_q u - \lambda\right).$$

We can extract the cases with $(X_i) = (n_i)$ by computing

$$G_{pq}^{(n_i)}(u, 0) = G_{pq}^{(n_i)}(u, (t_i))|_{(t_i)=0} = \left(\prod_i \partial_i^{n_i}\right) G_{pq}(u, (t_i))|_{(t_i)=0}$$

where ∂_i denotes differentiation by t_i. This can be done one t_i at a time in order to keep the computational complexity down. The conditional probability generating function is then given by

$$G_{pq}(u|(X_i) = (n_i)) = \sum_r \mathrm{P}\left[X_{pq} = r|(n_i)\right] u^r = \mathrm{E}\left[u^{X_{pq}}|(X_i) = (n_i)\right] = \frac{G_{pq}^{(n_i)}(u, 0)}{G_{pq}^{(n_i)}(1, 0)} \tag{6}$$

{eq:gf-cond}

where the conditional probabilities can be read of as the u coefficients. If I is big or the degrees high, however, this may require considerable computational time and space, and is therefore not practical in use. Still, it allows a comparison of the approximations with the exact distribution in a fair range of cases.

B.1. Exact distribution in the uniform case

The case with general λ_{ij} values will tend to be computationally demanding when $|I|$ or the n_i are large. However, the special case when the λ_{ij} are constant, the computations can be simplified. Recall that, for the distribution conditional on the marginal sums (n_i) is the same for all $(\lambda_{ij}) = (\alpha_i \alpha_j)$, and so we may choose $\lambda_{ij} = 1$.

For simplicity, we assume without loss of generality that $I = \{1, \ldots, m\}$ where $p = 1$ and $q = 2$. We can then write

$$G_{12}(u, (t_1, \ldots, t_m)) = H(u, t_1, \ldots, t_m) \cdot e^{-m(m-1)/2}, \quad H(u, t_1, \ldots, t_m) = \exp[f_m(u)]$$

where $\lambda = m(m-1)/2$ is the sum of all the $\lambda_{ij} = 1$ values and, suppressing the variables $t_1, \ldots, t_m$ in the notation, we write

$$f_k(u) = ut_1t_2 + \sum_{\substack{1 \le i < j \le k \\ (i,j) \ne (1,2)}} t_i t_j = f_{k-1}(u) + (t_1 + \cdots + t_{k-1})t_k, \quad f_2(u) = ut_1t_2.$$

Evaluating (6) requires computing the partial derivatives of $H(u)$ at $(t_i) = 0$, which we denote

$$H(u, t_1, \ldots, t_k | n_{k+1}, \ldots, n_m) = \partial_{k+1}^{n_{k+1}} \cdots \partial_m^{n_m} H(u, t_1, \ldots, t_m)|_{t_{k+1} = \cdots = t_m = 0}.$$

The critical observation is that we can write

$$H(u, t_1, \ldots, t_k | n_{k+1}, \ldots, n_m) = h_k(t_1 + \cdots + t_k) \cdot e^{f_k(u)}$$

where the dependency of $h_k(\phi)$ on $n_{k+1}, \ldots, n_m$ is implicit. This makes

$$\begin{aligned} H(u, t_1, \ldots, t_{k-1} | n_k, \ldots, n_m) =& \partial_k^{n_k} H(u, t_1, \ldots, t_k | n_{k+1}, \ldots, n_m)|_{t_k = 0} \\ =& \partial_k^{n_k} h_k(t_1 + \cdots + t_k) \cdot e^{f_k(u)}|_{t_k = 0} \\ =& \sum_{r=0}^{n_k} \binom{n_k}{r} h_k^{(r)}(t_1 + \cdots + t_k) \cdot (t_1 + \cdots + t_{k-1})^{n_k - r} \cdot e^{f_k(u)}|_{t_k = 0} \\ =& \sum_{r=0}^{n_k} \binom{n_k}{r} h_k^{(r)}(t_1 + \cdots + t_{k-1}) \cdot (t_1 + \cdots + t_{k-1})^{n_k - r} \cdot e^{f_{k-1}(u)} \end{aligned}$$

where $h_k^{(r)}(\phi)$ denotes the r-th derivative, so that for $k > 2$ we have

$$h_{k-1}(\phi) = \sum_{r=0}^{n_k} \binom{n_k}{r} h_k^{(r)}(\phi) \cdot \phi^{n_k - r}.$$

This makes $h_2(\phi)$ computable even for large m and n_i, providing us with

$$H(u, t_1, t_2 | n_3, \ldots, n_m) = h_2(t_1 + t_2) \cdot e^{ut_1t_2}$$

from which it follows that

$$H(u | n_1, \ldots, n_m) = \sum_{r=0}^{\min(n_1, n_2)} \frac{n_1!\, n_2!\, h_2^{(n_1 + n_2 - 2r)}(0)}{(n_1 - r)!\, (n_2 - r)!\, r!} \cdot u^r$$

Equation (6) then becomes

$$\sum_r \mathrm{P}\left[X_{pq} = r | (n_i)\right] u^r = \frac{G_{pq}^{(n_i)}(u, 0)}{G_{pq}^{(n_i)}(1, 0)} = \frac{H(u | n_1, \ldots, n_m)}{H(1 | n_1, \ldots, n_m)}.$$

References

[1] Guoliang Li, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, Chia-Lin Wei, Yijun Ruan, and Wing-Kin Sung. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, 11(2):R22, January 2010.