Appendix

A. The Boltzmann distribution

Two derivations of the Boltzmann distribution (taken from Atkins 2000, and Atkins 1978, respectively) are shown, the latter of which relates to the logic used by Berg and von Hippel for their derivation of the TF mismatch energies.

A.1 Derivation via the dominating energy distribution

Consider a system of N independent molecules (e.g. an ideal gas) with constant total energy E. While the amount of energy associated with any given particle cannot be determined one can address the question on how the energy is distributed over the particles on average. To this end it is useful to remember that energy levels are quantitized, that is, every molecule can occupy one of the available energy levels ε_0, ε_1, ε_2, …, ε_E where ε_0 is the level with lowest energy (arbitrarily set to a value of 0). At a given moment the system will be in a particular configuration where there will be n_0 molecules occupying energy level ε_0, n_1 molecules occupying level ε_1 and so forth. The occupation numbers n_i thereby change constantly due to the collision between the molecules, however, the total energy, E of the system and the total number of particles, N must stay unchanged:

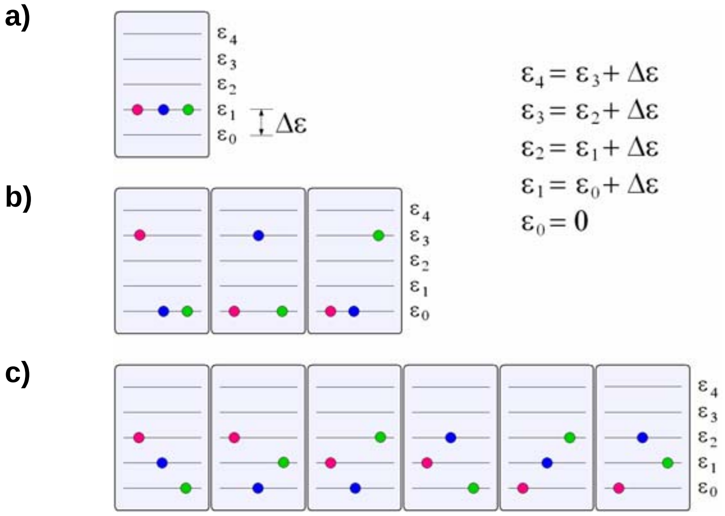
$$\sum_i \varepsilon_i n_i = E \tag{A.1}$$

and

$$\sum_i n_i = N\,. \tag{A.2}$$

To derive how many molecules will be associated with a given energy level on average we can start by considering a simple example. Suppose we have a system with five equally spaced energy levels which are given by $\varepsilon_i = i\ \Delta\varepsilon$ with $i \in \{0, 1, 2, 3, 4\}$. The system contains three particles with a total energy $E = 3\ \Delta\varepsilon$. We are interested in how the energy will distribute itself among the 3 particles, or in other words, how the particles will arrange themselves on the available energy levels. In one configuration the energy

Figure A.1 – Particle distribution across energy levels


a)
b)
c)
ε_4
ε_3
ε_2
ε_1
ε_0
$\Delta\varepsilon$
$\varepsilon_4 = \varepsilon_3 + \Delta\varepsilon$
$\varepsilon_3 = \varepsilon_2 + \Delta\varepsilon$
$\varepsilon_2 = \varepsilon_1 + \Delta\varepsilon$
$\varepsilon_1 = \varepsilon_0 + \Delta\varepsilon$
$\varepsilon_0 = 0$

a) There is only one way to distribute a total energy of 3 $\Delta\varepsilon$ evenly among the red, green and blue molecules. b) In comparison, there are three ways to assign the total energy to one of the three molecules. c) Finally, there are six ways to distribute the energy in such a way that one molecule has energy $\Delta\varepsilon$, one molecule has energy 2 $\Delta\varepsilon$ and one molecule has energy 0.

of the system is evenly distributed among the particles with each particle residing on energy level ε_1. As Figure A.1a illustrates, there is only one way in which this configuration or microstate of the system can be realized. Alternatively, the energy might be distributed in such a way that one particle carries all the energy and thus occupies energy level ε_3 while the other two particles have energy ε_0. As shown in Figure A.1b there are three ways in which this microstate can be attained. It is thus 3 times more likely than the first microstate. In a last set of configurations, the energy is distributed in such a way that one particle occupies energy level ε_0, one particle resides on energy level ε_1 and one particle occupies ε_2. There are six ways to achieve this energy distribution. This is thus the dominant configuration of the system, which will be found 60% of the time.

In general, the number of ways in which a configuration can be achieved is called the weight *W* of the configuration and is directly proportional to the probability of finding the system in a certain configuration. It is given by the binomial coefficient:

$$W = \frac{N!}{n_0!n_1!n_2!...n_E!} \tag{A.3}$$

where the n_i's are the occupation numbers for the different energy levels. In agreement with Figure A.1, the number of ways in which the first, second and third configuration in the above example can be realized is thus:

3! / (0! 3! 0! 0!) = 1,
3! / (2! 0! 0! 1! 0!) = 3,
3! / (1! 1! 1! 0! 0!) = 6.

Importantly, the larger the number of particles in a system the more dominant will one of the configurations be. In the following we thus seek to find this most dominant configuration that is, the n_i's that maximize W. For this we will rely on the case that N grow towards infinity (which is the case for most real world examples such as a mol of gas with ~6×10^{23} molecules). Dealing with ln(W) is thereby mathematically easier and yields identical results. To start with, imagine that changing the number of molecules on each energy level ε_i by the quantity dn_i goes in hand with a corresponding change dln(W) in the weight of the configuration, where dln(W) is:

$$d\ln(W) = dn_i \sum_i \frac{\delta \ln W}{\delta n_i}. \tag{A.4}$$

The maximum weight for the system can then be derived by setting $\mathrm{d}\ln(W) = 0$. Since we deal with a physical system there are two constraints that need to be fulfilled when shifting dn_i molecules. For one, the number of molecules in the system has to be kept constant, that is:

$$\sum_i dn_i = 0. \tag{A.5}$$

Secondly, the total energy of the system has to stay unchanged and therefore,

$$\sum_i \varepsilon_i dn_i = 0. \tag{A.6}$$

These two constraints are incorporated in finding the most probable configuration by using the method of Lagrange multipliers. Each constraint is thereby multiplied by a constant and added to the main condition. Using *α* and -*β* as the multipliers we obtain:

$$d\ln(W) = \sum_i \left(\frac{\delta \ln W}{\delta n_i} \right) dn_i + \alpha \sum_i dn_i + \beta \sum_i \varepsilon_i dn_i = 0\,. \qquad \text{(A.3)}$$

Using Sterling's formula for approximating *N*! and subsequently solving the equation (details on how this is done are provided for instance in [Physical-chemistry, Atkins]) one finds that the n_i's of the most probable configuration are given by the Boltzmann distribution:

$$n_i = N \frac{e^{-\beta\varepsilon_i}}{\sum_j e^{-\beta\varepsilon_j}}\,. \qquad \text{(A.4)}$$

where the term $e^{-\beta\varepsilon_i}$ is referred to as the Boltzmann factor for the state *i*. Equivalently, the probability p_i of a given particle to reside on energy level ε_i is:

$$p_i = \frac{n_i}{N} = \frac{e^{-\beta\varepsilon_i}}{\sum_j e^{-\beta\varepsilon_j}} = \frac{e^{-\beta\varepsilon_i}}{Z} \qquad \text{(A.5)}$$

where *Z* is the so called molecular partition function (Atkins). The relative probability of a molecule to be on energy level ε_i rather than on level ε_j is given by the ratio of the Boltzmann factors:

$$\frac{p_i}{p_j} = \frac{n_i}{n_j} = \frac{e^{-\beta\varepsilon_i}}{e^{-\beta\varepsilon_j}} \qquad \text{(A.6)}$$

For thermodynamic systems it can be shown that the scaling constant *β* is equal to 1/*kT* where *k* is the Boltzmann constant and *T* is the absolute temperature.

It is important to realize that looking at the most probable configuration is extremely useful since the distribution of particles in a system with *N* → ∞ will at any time deviate only slightly from the p_i's derived above. To see this, image a system with four equidistant energy levels (Figure A.2). If we set $\beta\ \Delta\varepsilon = 1$ then the probabilities for finding a particle with energy ε_i can be computed by:

$$p_i = \frac{e^{-i}}{\sum_{j=0}^{3} e^{-j}}.$$

Using this equation we find p_0, p_1, p_2 and p_3 to be 0.644, 0.237, 0.087 and 0.032, respectively. Given $N = 1000$ the weight of the most probable configuration is thus:

$$W_{\max} = \frac{1000!}{644!\ 237!\ 87!\ 32!}.$$

That means in the most probable configuration there are 644 molecules with energy ε_0, 237 with energy ε_1, 87 with energy ε_2 and 32 with energy ε_3. Now imagine one particle with energy $\varepsilon_3 = 3\ \Delta\varepsilon$ is transferring its energy to 3 particles with energy $\varepsilon_0 = 0$. Each of these latter particles would now have energy $\varepsilon_1 = \Delta\varepsilon$ while the former particle would have $\varepsilon_0 = 0$. The occupation numbers would thus have changed to 642, 240, 87 and 31. The weight of the resulting configuration would be:

$$W = \frac{1000!}{642!\ 240!\ 87!\ 31!}.$$

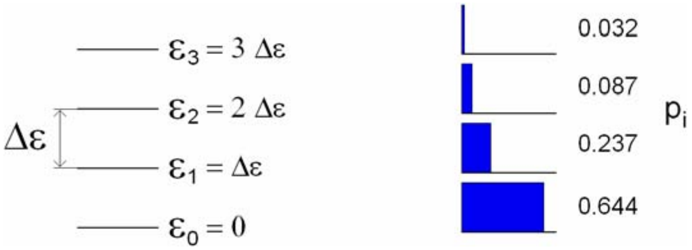
Dividing W by W_{max} yields a ratio of 0.97, that is, W is 0.97 times as likely as W_{max}. Changing the energy of 10 particles instead of one in the same fashion as above would result in a configuration with weight of:

$$W = \frac{1000!}{624!\ 267!\ 87!\ 22!}$$

where W / W_{max} would now be 0.028. Thus although the probabilities p_i would only be mildly effected (strongest for p_3 with going from 0.032 to 0.022) one would find the system only rarely in this configuration. The larger N the quicker this decay in the probabilities for other configurations takes place. For instance, using the same changes in the p_i's as above but with $N = 2000$ would yield a ratio W / W_{max} of 4.5×10^{-4}. In most real world examples N is of the order 6.0×10^{23} and a deviation of the p_i's from the values according to the most probable configuration is never observed. On the other hand, the

Boltzmann distribution is valid only if *N* is large. This can be seen for instance in the example of Figure A.1c where in the most likely configuration the particles are equally distributed over the energy levels and not according to the Boltzmann weights.

Figure A.2 – A system with four equidistant energy levels



$\varepsilon_3 = 3\ \Delta\varepsilon$
$\varepsilon_2 = 2\ \Delta\varepsilon$
$\Delta\varepsilon$
$\varepsilon_1 = \Delta\varepsilon$
$\varepsilon_0 = 0$
0.032
0.087
0.237
0.644
p_i

The left side shows the energy levels ε_0 through ε_3 for the theoretical system described in the text. The occupation probabilities given $\beta\,\Delta\varepsilon = 1$ are shown on the right.

A.2 Derivation via assuming contact with a heat reservoir

The Boltzmann distribution can be derived also by adopting a different point of view, which does not require writing down the weights of the different configurations of the system. This approach is similar to the one used by Berg and von Hippel for their derivation of the TF mismatch energies explained in Section 3.1.3 of the main text.

For the derivation of the Boltzmann distribution imagine a particle system as the one described above that is in contact with a large heat reservoir (which constitutes another particle system but of considerably larger size). The total energy of system and reservoir is E_{Tot}. If the energy of the system is E_i then the energy of the reservoir must be $E_{Tot} - E_i$. Let the number of ways, *W*, in which the reservoir can accommodate this energy be $W(E_{Tot} - E_i)$. Given the previous section we can easily imagine that the higher the energy of the reservoir the more configurations exist in which the reservoir can accommodate the energy (for instance, the energy $E = 0$ can be achieved by the reservoir only if all its molecules have energy $\varepsilon = 0$). Since we assume that the reservoir is much larger than the system we can assume that $E_{Tot} >> E_i$ virtually all the time. Therefore the number of configurations available to the reservoir with $E = E_{Tot} - E_i$ can be related to the number of configurations available at E_{Tot} via a Taylor expansion. If we again work with the logarithm of *W* then we can write:

$$\ln W\left(E_{Tot} - E_i\right) = \ln W\left(E_{Tot}\right) - E_i\left(\frac{\delta \ln W}{\delta E}\right)_{E_{Tot}} + \ldots \qquad \text{(A.8)}$$

where higher order terms of E_i can be neglected. The differential coefficient is dependent only on E_{Tot} and can therefore be written as a constant:

$$\beta = \left(\frac{\delta \ln W}{\delta E}\right)_{E_{Tot}} . \qquad \text{(A.9)}$$

(Notice the similarity between these expressions and the derivation by Berg and von Hippel for the TF mismatch energies outlined in the main text, page 43, where λ takes the role of β.) With this expression we obtain:

$$\ln W\left(E_{Tot} - E_i\right) = \ln W\left(E_{Tot}\right) - E_i\beta \;\rightarrow\; W\left(E_{Tot} - E_i\right) = W\left(E_{Tot}\right)e^{-\beta E_i} . \qquad \text{(A.10)}$$

Based on the discussion in Section A.1 one can reason that the probability, P_i, of the system having energy E_i is proportional to the number of ways the reservoir can accommodate the energy E_{Tot} – E_i in respect to the number of ways the reservoir can accommodate E_{Tot}. Therefore:

$$P\left(E_i\right) = C\frac{W\left(E_{Tot}\right)e^{-\beta E_i}}{W\left(E_{Tot}\right)} = Ce^{-\beta E_i} \qquad \text{(A.11)}$$

where C is a scaling constant. With the condition that the P_i's have to sum to 1, that is:

$$\sum_i Ce^{-\beta E_i} = 1 \;\rightarrow\; C = \frac{1}{\sum_i e^{-\beta E_i}} \qquad \text{(A.12)}$$

we obtain again the Boltzmann distribution:

$$P\left(E_i\right) = Ce^{-\beta E_i} = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} = \frac{e^{-\beta E_i}}{Z} \qquad \text{(A.13)}$$

this time for the probability of a system (which may consist of only one particle in which case we return to the molecular partition function) having energy E_i. Z is hereby referred to as the canonical partition function. Via the comparison to classical thermodynamics β can again be identified as being $1/kT$.

B. Zusammenfassung

Transkriptionsfaktoren (TFs) bilden Schlüsselkomponenten zellulärer regulatorischer Netzwerke, indem sie die Expression sowohl zelltypspezifischer als auch breit exprimierter Gene regulieren. Die Interaktion zwischen den Aminosäuren des jeweiligen Faktors und der DNA bildet die Grundlage für das sequenzspezifische Bindeverhalten der TFs, wobei ein gegebener Faktor eine Vielzahl von unterschiedlichen DNA Sequenzen binden kann, allerdings mit abweichender Affinität. Die Vielfältigkeit der Bindemuster und die enorme Länge eukaryotischer Genome machen die Vorhersage des Bindeverhaltens der TFs zu einem schwierigen Unterfangen. Traditionelle Methoden versuchen das Problem zu lösen, indem sie eine Unterteilung des Sequenzraums in Bindestellen und nicht gebundene Stellen vornehmen. Daß solche Modelle eine starke Vereinfachung darstellen, wird nicht zuletzt durch genomeweite Bindedaten belegt, die ein kontinuierliches Bindeverhalten von TFs aufzeigen.

Der erste Teil dieser Dissertation widmet sich deshalb der Entwicklung eines biophysikalischen Modells (genannt TRAP), das eine binäre Unterteilung zwischen Bindestellen (Hits) und ungebundenen Stellen vermeidet und stattdessen hoch und niedrig affine Sequenzen berücksichtigt. Wie gezeigt wird, können die Parameter des Modells durch eine physikalisch motivierte Vorschrift bestimmt werden, die für alle untersuchten Organismen von Hefe bis zu Mensch gilt. Die konzeptionelle, sowie praktische Überlegenheit von TRAP gegenüber traditionellen Hit-basierten, sowie alternativen affinitätsbezogenen Methoden, wird dargestellt.

Um TFs zu detektieren, die für die Regulation ganzer Gengruppen verantwortlich sind, wurde TRAP im Folgenden durch ein statistisches Verfahren erweitert, das mittels einer Reihe hypergeometrischer Tests prüft, ob eine Anreicherung potentieller Zielgene eines gegebenen TFs innerhalb einer benutzerdefinierten Gengruppe existiert. Die Anwendung dieser Methode (genannt PASTAA) auf Gruppen gewebespezifischer Gene ermöglichte die Identifizierung einer umfassenden Anzahl experimentell bekannter TF-Gewebe-Assoziationen. PASTAA war hierbei erheblich erfolgreicher als verschiedene alternative Methoden. Darüber hinaus ließen die Resultate eine Reihe interessanter, biologischer Schlussfolgerungen zu, wie z.B., daß hochaffine Bindestellen gewebespeziefischer TFs bevorzugt in proximalen Promotoren, upstream vom Transkriptionsstart vorkommen. Die Analyse war dabei robust gegenüber der Auswahl an Promotersequenzen und der Herkunft der Expressionsdaten.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, April 2008   Helge G. Roider