ABSTRACT

Motivation: One of the central questions in the post-genomic era is the understanding of protein–protein interactions and of protein complex formation. It has been observed that protein complex size distributions of the yeast Saccharomyces cerevisiae decay exponentially. The shape of these size distributions reflects mechanisms of protein complex association and dissociation.

Results: We present the most simple dynamic model that is able to reproduce the observed protein complex size distribution for yeast. This protein association–dissociation model (PAD-model) simulates the dynamics of protein complex formation on a genomic scale for about 50 million protein molecules. By ruling out different model variants it is possible to elucidate fundamental features of the protein complex dynamics, e.g. complex association is independent of complex size. In addition, the PAD-model provides information about the complexity of the yeast proteome and it gives an idea of how many complexes could not be identified during the measurements.

Availability: All programs used for this publication are available on request from the authors.

Contact: beyer@imb-jena.de

Supplementary information: Supplementary information about the model and its interpretation can be downloaded from http://www.imb-jena.de/tsb/pad

INTRODUCTION

Understanding the dynamics and complexity of the proteome is one of the core challenges of the post-genomic era. The dynamics of protein interactions is central for whole cell functioning (Ideker et al., 2001; Hlavacek et al., 2003). Simulations of the entire cell, performed to elucidate emergent statistical and dynamic properties, need detailed knowledge of the proteome dynamics (Endy and Brent, 2001; Tomita, 2001). Most cellular functions are conducted or regulated by protein complexes of varying size (Alberts et al., 2002). It is mainly the proteome (i.e. the entirety of all expressed proteins in an organism and their interactions) that actually "executes" the genetic code. The genome size of different organisms is much more similar than the difference in complexity of the organisms implies (the apparent paradox of "too few" genes in complex organisms (Rubin, 2001)). Hence, it has been suggested that protein complexes may contribute substantially to an organism’s complexity (International Human Genome Sequence Consortium, 2001; Venter et al., 2001; Gavin et al., 2002; Hollunder et al., 2005). An organism with 6000 different proteins can hypothetically create about 18 million different pairs of interacting proteins. Assuming complexes of size 3 we already get about $10^{13}$ different variants. This combinatorial explosion (Endy and Brent, 2001; Hlavacek et al., 2003) shows how evolution could significantly increase the regulatory and metabolic complexity of organisms without substantially increasing the genome size (Koonin et al., 2002; Gavin and Superti-Furga, 2003). However, only a very small subset of these many possible protein complexes is actually realized (Sear, 2004).

Recent large-scale measurements of characteristics of the yeast proteome (Gavin et al., 2002; Ho et al., 2002; Ghaemmaghami et al., 2003; Hurowitz and Brown, 2004; Beyer et al., 2004) paved the way for performing integrated analyses of the processes underlying such data, thereby improving our understanding of the proteome as a whole. Studies of this kind yield substantial new insights into cell regulation and cell dynamics (Ideker et al., 2001; Beyer et al., 2004). Based on large-scale measurements of protein complexes in the yeast Saccharomyces cerevisiae it was shown that the size-frequency distributions of such complexes have a number of common characteristics (Wilhelm et al., 2003). When plotting the number of different complexes of a given size versus the complex size one obtains an exponentially decreasing distribution. The authors hypothesize that the shape of this distribution reflects the nature of the underlying cellular dynamics which is creating the protein complexes. A simulation model that correctly depicts the most important processes should therefore yield the same complex size distribution.

Simulating the interaction of a large number of components is a well established technique in bioinformatics and systems biology that has successfully been applied in the past (Barabási et al., 1999; Endy and Brent, 2001; Dittrich et al., 2001; Rzhetsky and Gomez, 2001; Koonin et al., 2002; Furusawa and Kaneko, 2003; Sear, 2004). So-called birth, death and innovation models (BDIMs) have been used to reconstruct protein domain frequency distributions and to draw non-trivial conclusions (Wolf et al., 1999; Qian et al., 2001; Koonin et al., 2002; Karev et al., 2004). Our strategy follows a similar line of thought: if the model is capable of reproducing the observed protein complex frequency distribution, it can be used to draw instructive conclusions.

Here we outline the most simple dynamic model that explicitly simulates association and dissociation of protein complexes and that is capable of reproducing the observed complex size distribution in S. cerevisiae. The model that we propose simulates only two processes: association and dissociation of protein complexes. During discrete time steps single proteins and protein complexes are randomly selected to undergo association and dissociation. The
system is considered to be spatially homogeneous. The value of such a simplistic model is that it helps in identifying the dynamics of the most fundamental processes underlying protein complex formation. This model allows us to answer the question as to which processes and what kind of dynamics are needed at minimum in order to reproduce the observed distribution. By comparing different variants of such a model it is possible to rule out certain mechanisms of protein complex formation and the relative importance of model parameters can be estimated. In addition, the model opens up the possibility (to our knowledge for the first time) of assessing the complexity of the yeast proteome as a whole. This gives us some idea of how much has been missed during the protein complex measurements.

**METHODS**

**Protein complex data**

The system that we simulate corresponds to an entire hypothetical yeast cell. We have analyzed protein complex data by Gavin et al. (2002), which were obtained by tandem affinity purification (TAP) and subsequent mass-spectrometric protein identification. Here we use the manually curated TAP complexes (Gavin et al., 2002), which is a set of 232 biologically meaningful complexes in yeast with sizes ranging from 1 to 88 different proteins per complex. In this study we only account for complexes containing at least two different proteins, yielding a subset of 229 complexes. For this data set, we calculated the complex size-frequency distribution (Fig. 1), which shows that the number of different complexes of a given size decays nearly exponentially.

**Protein abundance data**

For simulating complex formation we need data on the abundance of the ~6200 different proteins in yeast cells. Since we simulate hypothetical proteins we do not need actual concentrations for every yeast protein, but we only need a realistic protein abundance distribution for parameterizing the model. The protein abundance distribution that we use follows protein concentrations compiled by Beyer et al. (2004) in the upper range with data from Ghaemmaghami et al. (2003) and Greenbaum et al. (2003). The data by Beyer et al. (2004) may underestimate the number of low-abundance proteins, because smaller signals are more difficult to detect. We, therefore, assume that 2000 proteins not contained in the Beyer et al. dataset are mainly expressed at low concentrations. The resulting abundance distribution roughly follows a power-law. More details about the protein abundance distribution are given in the supplementary information.

**Dynamic complex formation model**

We discuss three variants of the protein complex association–dissociation model (PAD-model) with the following features:

1. In all three versions the composition of the proteome (i.e. the abundance of all proteins) does not change with time. This means, degradation of proteins is always balanced by an equal production of the same kind of proteins. This assumption is justified by the fact that we simulate the averaged complex size distribution of a large number of cells under constant log-growth conditions (Gavin et al., 2002).

2. The cell consists of either one (PAD-A and B) or several (PAD-C) compartments in which proteins and protein complexes can freely interact with each other. Thus, all proteins can potentially bind to all other proteins in their compartments.

3. Association and dissociation rate constants are the same for all proteins (i.e. the values chosen represent cellular averages). In PAD-models A and C association and dissociation are independent of complex size and complex structure.

4. At each time step, a set of complexes is randomly selected to undergo association and dissociation. Association is simulated as the creation of new complexes by the binding of two smaller complexes and dissociation is simulated as the reverse process, i.e. it is the decay of a complex into two smaller complexes. The number of associations and dissociations per time step is $k_i \cdot N_i^2$ and $k_i \cdot N_i$, respectively, with $N_i$ being the total number of complexes in the cell and $k_i(1/\text{no. of complexes} \times \text{time})$ being the association and dissociation rate constants, respectively. The rate constants $k_i$ and $k_j$ are mathematically equivalent to biochemical rates of a reversible reaction (see Supplementary information).

**Fig. 1.** Protein complex size distribution. Cumulative frequency (starting from the right) of different protein complexes of the given size. Complex size is measured as number of different proteins in a complex (measurable complex size). Frequency refers to number of distinct complexes of the given size. Diamonds: measured complex frequencies of 229 TAP measured complexes (Gavin et al., 2002). Solid lines: Best fit of dynamic simulation at steady-state using PAD-model A (no partitioning) and PAD-model C (with partitioning). Solid lines are averaged results of several model runs where 229 distinct complexes were randomly selected from a pool of simulated complexes. Dashed line: exponential fit. This figure can be viewed in colour on Bioinformatics online.

PAD-A is the simplest model where all proteins can interact with each other (no partitioning) and it assumes that association and dissociation are independent of complex size. PAD-B is equivalent to PAD-A, except for the assumption that larger complexes are more likely to bind (preferential attachment). In this case we assume that the binding probability is proportional to $i \cdot j$, where $i$ and $j$ are the sizes of two potentially interacting complexes.

Finally, model PAD-C extends PAD-A by assuming that proteins can interact only within groups of proteins (with partitioning). The sizes of these protein groups are based on the sizes of first level functional modules according to the yeast database CYGD (http://mips.gsf.de). PAD-C assumes 16 modules, each containing between 100 and 1000 different ORFs. Hence, the protein groups do not represent physical compartments, but rather resemble functional modules of interacting proteins. The total steady-state pool of protein complexes is obtained by independently simulating each protein group, which gives rise to one ‘complex pool’ per group. In order to get whole-cell averages, results are averaged based on the sizes of the resulting complex pools. For instance, when averaging the resulting bait distributions, it is assumed that complexes are drawn with higher probability from larger pools.

**Mathematical description**

Since explicit simulation of an entire cell (we simulate ~50 million protein molecules) is too time-consuming for many applications of the model, we also developed a mathematical description of the PAD model, which allows us to assess different scenarios and parameter combinations more quickly.
The change of the number of complexes of size \( i \), \( \Delta x_i \), during one time step \( \Delta t \) can be described as

\[
\frac{\Delta x_i}{\Delta t} = G^a_i + G^d_i - L^a_i - L^d_i,
\]

(1)

where \( G^a \) and \( G^d \) are the gains due to association and dissociation and \( L^a \) and \( L^d \) are losses due to association and dissociation, respectively. \( G^a \) is the total number of associations of complexes smaller than \( i \) that together form complexes of size \( i \). When complexes larger than \( i \) dissociate they potentially create fragments of size \( i \). The gain \( G^d \) is the number of these fragments. For PAD-A the term for association can be computed as (see Supplementary information)

\[
G^a_i = L^2_i = k_a \cdot N_C \left[ \sum_{j=1}^{i-1} x_j \cdot x_{i-j} - x_i \right] / \left( N_C - 1 \right).
\]

(2)

The dissociation term reads

\[
G^d_i = L^4_i = k_d \cdot \sum_{j=2}^{i-1} \left[ \sum_{j=1}^{i-1} x_j \cdot \left( j \cdot x_j / x_i \right) \right] N_j - x_i.
\]

(3)

where \( k_a \) and \( k_d \) are the respective association and dissociation rate constants, \( x_i \) is the number of complexes of size \( i \), \( x_i \) is a correction term for even complexes sizes \( i \), \( N_C = \sum x_i \) is the total number of protein complexes and \( N_j = 2^{(j-1)} - 1 \) is the number of possible dissociations of a complex of size \( j \).

Figure S1 (Supplementary information) shows a comparison of a numerical solution of Equation (1) with a stochastic simulation of the association–dissociation process. After a transient period a steady-state is reached. We are mainly interested in this steady-state distribution of frequencies \( x_i \). Hence, we have to find a set of \( x_i \) solving \( \Delta x_i / \Delta t = 0 \). The solution of this non-linear equation system is obtained by numerically minimizing all \( \Delta x_i / \Delta t \). Dividing Equation (1) by \( k_a \), it can be seen that the steady-state distribution is independent of the absolute values of \( k_a \) and \( k_d \), but it only depends on the ratio of the two parameters \( R_{ad} = k_a / k_d \). Hence, only two parameters affect \( x_i \) at steady-state: the total number of proteins \( N_P \) (which indirectly determines \( N_C \)) and the ratio of the two rate constants \( R_{ad} \).

The dissociation terms remain unchanged for PAD-B model, whereas the association terms have to be modified (see Supplementary information). In case of PAD-C we calculated weighted averages of results obtained with PAD-A.

**Measurable size distribution and bait selection**

Based on the distribution resulting from Equation (1) at steady-state, we derive two further distributions: (1) the ‘measurable size distribution’ and (2) the ‘bait distribution’. The former is defined as the frequency distribution of the measurable complex sizes. The measurable complex size is the number of different proteins in a protein complex (as opposed to the total number of proteins). For the measurable size distribution we only count the number of complexes with distinct protein compositions (Fig. 2). In order to determine the distribution of measurable complex sizes corresponding to the steady-state distribution, we create a set of complexes according to the original steady-state size distribution by randomly ‘filling’ the complexes with proteins from the protein abundance distribution. We then compute the resulting measurable size distribution. Results shown are the averages of several of such random sets.

The bait distribution is used to compare the simulation to the actual TAP measurements (Gavin et al., 2002). The bait distribution is obtained by randomly selecting and subsequently analyzing a subset from all simulated complexes. We call that distribution bait distribution, because the process of selecting a subset of all complexes corresponds to selecting bait proteins for pulling out the complexes during the measurements (Gavin et al., 2002). We always select 229 different complexes, which is the number of TAP complexes with which we compare the simulations.

**Fig. 2.** Original and measurable complex size distributions. The measurable complex size refers to the number of different proteins in a protein complex. In addition, for the measurable complex frequency only distinct complexes get counted. Thus, there are two differences in comparison to the original size distribution: the distributions differ with respect to the counting of proteins and with respect to the counting of complexes. (a) Examples of different complexes, which all have the measurable complex size 2. Note that complexes A and B are not counted as different complexes, because they contain proteins of the same type. Hence, for the measurable complex size distribution the three complexes would be counted as two complexes of size 2. (b) Measurable versus actual complex size distribution. Diamonds show frequencies of measurable complexes. Filled diamonds and triangles reflect simulation without partitioning (PAD-A) and open diamonds and triangles are simulation results assuming binding only within certain modules (PAD-C). The difference between the original and the measurable complex size distribution is comparably small, because most of the simulated complexes are unique. However, in case of PAD-C smaller complexes occur at higher copy numbers and larger complexes are often counted as smaller measurable complexes because they contain some proteins more than once. This figure can be viewed in colour on Bioinformatics online.

**Computation of a dissociation constant \( K_D \)**

Mathematically, our model describes a reversible (bio-)chemical reaction (see Supplementary information). When following this interpretation of the model one can calculate an equilibrium dissociation constant \( K_D \), which quantifies the fraction of free subcomplexes A and B compared with the bound complex AB. Interestingly, this equilibrium is complex size dependent, because a large complex AB is less likely to randomly dissociate exactly into the two specific subunits A and B than a small complex. (Note that A and B can be ensembles of several proteins.) Therefore, we get for any given complex of size \( i \) the following \( K_D \):

\[
K_D(i) = [A][B]/[AB] = (R_{ad} \cdot N_i \cdot V)^{-1},
\]

(4)

where \( N_i \) is the number of possible fragments of a complex of size \( i \) and \( V \) is the cell volume. Cell-wide averages of \( K_D \)-values are estimated by computing a weighted average \( \bar{K}_D = \sum_i K_D(i) \cdot x_i / N_C \), with \( N_C \) being the total number of complexes and \( x_i \) being the number of complexes of size \( i \).
Quality of regression

The quality of the regressions is quantified by the sum of squared and weighted residuals $\Delta X$:

$$\Delta X = \sum_{i=2}^{i_{\text{max}}} (X_i - \bar{X}_i)^2 / X_i^2,$$  \hspace{1cm} (5)

where $X_i$ is the cumulative frequency of complexes down to size $i$ of the measured bait distribution and $\bar{X}_i$ is the corresponding cumulative frequency of the simulated bait distribution. Frequencies are cumulated starting with the largest measured complex (size 88) and counting down to size $i$ (Fig. 1).

RESULTS

We dynamically simulated the association and dissociation of 6200 different protein types yielding a set of about 50 million protein molecules. Subsequently we analyzed the resulting steady-state size distribution of protein complexes. This steady-state is thought to reflect the log-growth conditions under which the yeast cells were held when TAP-measuring the protein complexes (Gavin et al., 2002). Based on measured protein complex data (Gavin et al., 2002) we calculated a protein complex size distribution which we can compare the simulation results (Fig. 1).

The TAP measurements do not provide concentrations of the measured complexes, but they only demonstrate the presence of a certain protein complex in yeast cells. In addition, the number of proteins of a certain type inside such a complex could not be measured. Hence, the complex size from Figure 1 does not represent real complex sizes (i.e. total number of proteins in the complex), but it refers to the number of different proteins in a complex. The measured data reflect the characteristics of only 229 different protein complexes of size 2 and larger, which is just a small subset of the ‘complexosome’. These peculiarities have to be taken into account when comparing simulation results to the observed complex size distribution. We refer to the ‘measurable complex size’ as the number of distinct proteins in a protein complex (Fig. 2). When comparing our simulation results with the measurements, we always select a random subset of 229 different complexes from the simulated pool of complexes. This results in a complex size distribution comparable to the measured distribution from Figure 1 (bait distribution).

An exponential fit ($a \cdot e^{bi}; i$ is the complex size) of the observed complex size distribution yields a slope $b = -0.063$, which is slightly shallower than the slopes obtained for other datasets (Wilhelm et al., 2003). In agreement with previous findings the TAP complexes follow an exponential-law much more than a power-law (Wilhelm et al., 2003). For instance, the residuals $\Delta X$ between the observed size distribution and the regression are 0.68 for the exponential fit and 5.0 for the respective power-law regression. Also the $R^2$ deviate significantly (exponential fit: 0.98, power-law fit: 0.92). The best fit simulation using the simplest model PAD-A yields a $\Delta X$ of 0.75, which is slightly above the exponential fit. The complex size distributions resulting from simulations with PAD-A are close to an exponential distribution. Koonin et al. (2002) used a mathematically similar model to simulate the evolution of protein–protein interactions. They also found that the number of distinct domains in proteins decreases exponentially, whereas the total number of domains per protein follows a power-law. In our model the former roughly corresponds to the measurable size distribution and the latter to the ‘original’ size distribution. The original size distribution of PAD-A is slightly closer to a power-law than the measurable size distribution (Fig. 2), which is consistent with the findings by Koonin et al. (2002).

The dynamic simulation needs only very few input parameters: a set of proteins, which is defined by the protein abundance distribution, and the association and dissociation rate constants, while only the ratio of the two rate constants $R_{ad}$ actually matters (see Methods). Since the protein abundance distribution is known (Beyer et al., 2004), the ratio of $R_{ad}$ is the only ‘free’ (i.e. adjustable) parameter of PAD-A. The best fit of the model to the observed complex size distribution is obtained for a $R_{ad}$ of $1.6 \times 10^{-7}$. Using Equation (4) and assuming a cell volume of $5 \times 10^{-11}$ ml one gets $K_{D}$-values ranging between $10^5$ (size 2) and $10^{-25}$ nM (size 100) for this $R_{ad}$.

A comparison of the subset of 229 complexes (Fig. 1) and the complete simulated complexosome (Fig. 2) reveals a significant deviation of the two distributions. This indicates that only a small subset of all protein complexes in the cell has been observed during the TAP measurements. However, the measurable size distribution closely follows the distribution of the ‘actual’ complex sizes in the simulated pool. Hence, only few complexes contain a protein more than once and hardly any complexes (except for very small ones) occur more than once in the simulated system. The simulated cell contains more than 3.5 million different complexes, which certainly exceeds the true number of different complexes in yeast cells. A crucial reason for this overestimation is that specificity of protein binding is not taken into account in PAD-A (see Discussion).

Since the model assumes that association and dissociation are independent of the complexes’ composition, the number of the different protein types (i.e. number of expressed genes) has no effect on the distribution of actual complex sizes. Only the measurable complex size, i.e. the number of different proteins in a complex, is affected by the composition of the protein pool. Yet even this effect can be reduced to one characteristic value of the protein abundance-distribution. The distribution of measurable complex sizes is mainly determined by the probability $P_{ad}$ that two randomly chosen molecules from the proteome are identical (see Supplementary information). In other words, all protein abundance distributions having the same probability $P_{ad}$ yield very similar complex size distributions. This reduces the relevant model parameters to three input parameters: $R_{ad}$, $N_p$ and $P_{ad}$. Among the model parameters, the kinetic parameter $R_{ad}$ and the number of proteins $N_p$ are the most sensitive ones (Table 1). Already a 10% deviation of $R_{ad}$ or $N_p$ from the best fit scenario may double the weighted residual $\Delta X$. However, none of the model parameters significantly changes the shape and slope of the distribution, i.e. the distributions are ‘close-to-exponential’ independent of the selected parameters.

No preferential attachment

A core assumption of our complex formation model PAD-A is that association and dissociation are independent of the complex size. One possible extension of this model is to assume that association of larger complexes is more likely than association of smaller complexes. The rationale for this preferential attachment might be that larger complexes have more potential interaction surfaces than smaller complexes. However, sterical effects might oppose such preferential attachment. In order to test the hypothesis that association might depend on complex size we simulated a second model that includes a size dependence of association (PAD-B). Since association of complexes is an undirected process we presume that both
binding partners equally contribute to the binding affinity. Hence, in the preferential attachment model two complexes associate with a likelihood proportional to the product of the complex sizes.

Figure 3 shows that the assumption of preferential attachment leads to a power-law complex size distribution. This result is in agreement with the previous finding that preferential attachment in the context of network growth leads to power-law connectivity distributions (Barabási et al., 1999). However, the experimentally observed distribution clearly follows an exponential-law [Fig. 1 and (Wilhelm et al., 2003)]. The fundamentally different character of these two distributions suggests that generally there is no preferential attachment during protein complex formation.

**Specific binding**

Another important simplification of PAD-A is that all proteins can potentially interact with each other. Obviously this is a vast simplification of biological reality. As a first step towards a better approximation of real protein–protein interactions we group the proteins into 16 modules of different sizes (corresponding to functional modules, see Methods section) and we restrict protein–protein interactions to proteins belonging to the same module (PAD-C). The best fit of PAD-C ($R_{ad} = 3.8 \times 10^{-6}$) yields better results than the best fit of PAD-A. In this case the weighted residual $\Delta X$ drops to 0.63, which is even below the $\Delta X$ of the exponential fit (Table 1). The character of the resulting distribution is the same as for PAD-A (i.e. close to exponential). However, the parameter for the best fit deviates significantly: The optimal $R_{ad}$ for PAD-C is larger by more than an order of magnitude (Table 1). In order to match the observations, the simulated affinity must be much larger if we assume a smaller number of interacting partners. Also, the number of different complexes drops significantly (PAD-A: 3.5 million, PAD-C: 2 million).

**DISCUSSION**

We developed a very simple, dynamic model that is capable of reproducing the observed complex size distribution. Given the small number of input parameters the very good fit of the observed data is astonishing. A number of conclusions with respect to the processes underlying protein complex formation can be drawn.

First, preferential attachment does not take place in yeast cells under the investigated conditions. The absence of preferential attachment is biologically plausible: specific and strong binding can be just as important for small protein complexes as for large complexes. This implies that also the dissociation should on average be independent of the complex size. The interpretation of the simulated association and dissociation in terms of $K_D$-values suggests that larger complexes bind more strongly than smaller complexes. [See the Supplementary information for a detailed discussion of Equation (4).] However, the size dependence of $K_D$ is compensated by the higher number of possible dissociations in larger complexes. In all variants of the PAD-model we assume that all possible dissociations happen with the same probability. In reality it is more plausible that large complexes break into specific subcomplexes, which subsequently can be reused for a different purpose (Gavin and Superti-Furga, 2003; Hollunder et al., 2005). An improved version of the model should not only account for the specificity of association, but also for specific dissociation.

### Table 1. Parameter sensitivity of the dynamic complex formation model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Residual $\Delta X^a$</th>
<th>Exponent$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential fit of measurements</td>
<td>0.68</td>
<td>-0.063</td>
</tr>
<tr>
<td>Power-law fit of measurements</td>
<td>4.98</td>
<td>-0.067</td>
</tr>
<tr>
<td>Best fit (PAD-model $A^c$)</td>
<td>0.75</td>
<td>-0.067</td>
</tr>
<tr>
<td>($R_{ad} = 1.6 \times 10^{-6}$, $N_{p} = 5 \times 10^{7}$, $P_{ad} = 0.0028$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best fit (PAD-model $C^c$)</td>
<td>0.63</td>
<td>-0.065</td>
</tr>
<tr>
<td>($R_{ad} = 3.8 \times 10^{-6}$, $N_{p} = 5 \times 10^{7}$, $P_{ad} = 0.0028$)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Sum of squared and weighted differences between observed and simulated cumulative frequencies (see Methods section).

$^b$ Exponent of exponential fit of measured and simulated cumulative frequency distribution. This parameter quantifies how the shapes of the distributions deviate from each other.

$^c$ Best fit: Simulation that best fits the observed complex size distribution using the mechanistic models PAD-A (no partitioning) or PAD-C (with partitioning). To obtain these fits only $R_{ad}$ was calibrated, while $N_{p}$ and $P_{ad}$ were derived from independent measurements.

$^d$ $R_{ad}$ up/down: association-to-dissociation ratio increased/reduced by 10%.

$^e$ $N_{p}$ up/down: total number of proteins increased/reduced by 10%.

$^f$ $P_{ad}$ up/down: probability of finding the same protein twice increased/reduced by 10%.

The parameter variations in rows 5–10 are changes with respect to the scenario ‘best fit (PAD-model $A^c$).’

![Fig. 3. Effect of preferential attachment. Cumulative number of distinct protein complexes versus their size, resulting from simulations without (diagonals) and with (squares) preferential attachment to larger complexes. Both simulations are performed with the best fit parameters for PAD-A. Results are shown for a subset of 229 bait complexes that have been selected and analyzed as in Figure 1. Solid lines are regressions through the simulated frequencies. In case of preferential attachment the best regression result is obtained with a power-law, while the simulation without preferential attachment is best fitted assuming an exponentially decreasing curve. Preferential attachment (PAD-B) is simulated assuming that the probability of an association is proportional to the product of the sizes of the two associating complexes. The original, measurable and bait distributions are always close to exponential in case of PAD-A and power-law like in case of PAD-B, independent of the parameters chosen. This figure can be viewed in colour on Bioinformatics online.]
A second important conclusion is that the number of complexes that were missed during the TAP measurements is potentially large. Based on our simulations, for the first time we can give an upper limit of the number of different complexes in cells. At first glance, the number of different complexes in PAD-A (>3.5 million) and PAD-C (~2 million) appear to be far too large. Even PAD-C may overestimate the true number of different complexes, because association within the groups is unrestricted. However, the PAD-models do not only simulate functional, mature complexes, but they also consider all intermediate steps. Each of these steps is counted as a different protein complex. The large difference between the number of measured complexes and the (potential) number of existing complexes may partly explain the very small overlap that has been observed between different large scale measurements of protein complexes (von Mering et al., 2002; Wilhelm et al., 2003).

A correct interpretation of the kinetic parameters is important. First of all, $k_a$ and $k_d$ cannot be compared to real numbers, because the model does not define a length of the time steps for interpreting $k_a$ and $k_d$ as actual rate constants. In addition, the association-to-dissociation ratio $R_{ad}$ is not identical to a physical $K_D$-value obtained by in vitro measurements of protein binding in water solutions. Several reasons do not allow for this simple interpretation:

1. In vivo diffusion rates are below those in water due to the high concentration of proteins and other large molecules in the cytosol (Endy and Brent, 2001).

2. Most proteins are either synthesized where they are needed or they get transported directly to the site where the complex gets compiled (Alberts et al., 2002). Hence, transport to the site of action is on average faster than random diffusion.

3. Protein concentrations are often above the cell average due to the compartmentalization of the cell. All these processes (protein production, transport and degradation) are not explicitly described in the PAD-model, but they are lumped in our assumptions. The $R_{ad}$ (and the $K_D$ derived from it) must, therefore, be interpreted as an operationally defined property. It characterizes the overall, cell averaged complex assembly process, which includes all steps necessary to synthesize a protein complex.

However, even the model-derived $K_D$-values allow for some conclusions regarding complex formation. We calculated weighted averages ($\bar{K}_D$) of the size-dependent $K_D$-values by using the steady-state complex size distribution of the best fit (see Methods). This yields average $\bar{K}_D$-values of 4.7 and 0.18 nM for the best fits of PAD-A and PAD-C, respectively. First, the fact that the $\bar{K}_D$ for PAD-C is below that of PAD-A underlines the notion that more specific binding is reflected by smaller $K_D$ values (Sear, 2004). Second, typical in vitro $K_D$-values are above 1 nM (Amini et al., 2003; Sear, 2004), thus the average $K_D$ of PAD-C is comparably low. The model, therefore, confirms that protein complex formation in vivo gets accelerated due to directed protein transport and due to the compartmentalization of eukaryotes. It is a surprising finding though, that important aspects of these highly regulated protein synthesis and transport processes can on average be described by a simple compartment model assuming random association and dissociation.

Large scale protein–protein interaction data sets are subject to substantial error (von Mering et al., 2002; Grünewald and Wintzeler, 2002), resulting in a potentially large number of false positives and false negatives. In order to get a correct picture of the protein complex size distribution it is necessary to have an unbiased, random subset of all complexes in the cells. It is known that there are certain biases in the TAP data, e.g. there is a reduced number of membrane proteins among the bait proteins (Gavin et al., 2002). However, if compared to other datasets such as MIPS complexes (http://mips.gsf.de), the TAP complexes constitute a fairly random selection of all protein complexes in yeast (von Mering et al., 2002). Uncertainties in the TAP data do not affect our conclusions as long as they are not strongly biased with respect to the resulting complex size distribution. Since Gavin et al. (2002) have measured long-term interactions, our results apply to permanent complexes. Yet the model is applicable to future protein complex data that take account of transient binding.

In this study we use the protein complex data from Gavin et al. (2002) as an example to outline the methodology. The model is equally applicable to any other suitable complex data set, such as the data from Ho et al. (2002). In particular, the model can be applied to more reliable data in the future and it is useful for analyzing protein complex formation in other organisms. It then becomes possible to compare $R_{ad}$-values (or $K_P$-values) of different organisms and of different cell types. For example, one question that could be addressed is, whether larger cells exhibit different $R_{ad}$-values. It is also interesting to see if the $R_{ad}$ is larger or smaller in quickly proliferating cells compared to quiescent cells. This could be helpful for better understanding and characterizing tumor cells.

The simulated complex size distribution is almost independent of the assumed protein abundance distribution. The detailed shape hardly matters and only $P_P$ (i.e. the probability of randomly selecting the same protein twice) has some influence (Table 1). It turns out that $P_P$ is a valuable summarizing property that can be used to characterize proteomes of different species. A decreasing $P_P$ increases the number of different large complexes (the slope in Table 1 gets more shallow) because it is less likely that a large complex contains the same protein twice. Thus, $P_P$ is a measure of complexity that not only relates to the diversity of the proteome but also to the composition of protein complexes.

Probably the most severe simplification in our model is the assumption that all proteins can potentially interact with each other. The PAD-model C is a first step towards more biological realism. By restricting the number of potential interaction partners it more closely maps functional modules and cell compartments, both of which restrict the interaction among proteins (Wilhelm et al., 2003). The partitioning in PAD-C connotes that proteins within one group exhibit very strong binding, whereas binding between protein groups is set to zero. This again is a simplification, since cross-talk between different modules or compartments is possible.

Future extensions of the model could incorporate more and more detailed information about the binding specificity of proteins. Assuming even more specific binding will further reduce the number of different complexes, the frequency of the complexes will increase. High-binding specificity potentially lowers the complex sizes, so $R_{ad}$ has to be increased in order to fit the experimentally observed protein complex size distribution. On the other hand, cross-talk gives rise to larger complexes. Taking both counteracting refinements into account, it is impossible to generally predict the best fit $R_{ad}$ since it depends on the quantitative details. A first additional refinement of PAD-C could account for the observed clustering of protein interaction networks (Wilhelm et al., 2003). In a second step one could simulate protein associations and dissociations according
to predefined binary protein interactions. A very detailed model could additionally account for individual association/dissociation rates between individual proteins.

Such extensions will yield more realistic figures about the number of different protein complexes created in yeast cells. However, starting model development with the most simple assumptions reveals the most important characteristics of the system for reproducing the observations. The excellent match that we have already obtained with the most simple model PAD-A is striking.

ACKNOWLEDGEMENT
This work has been funded by the Federal Ministry of Education and Research, Germany (grant 0312704E).

REFERENCES