# Optimal Reaction Coordinate as a Biomarker for the Dynamics of Recovery from Kidney Transplant

**Sergei V. Krivov[1,2]*, Hayley Fenton[3], Paul J. Goldsmith[4], Rajendra K. Prasad[4], Julie Fisher[2,3], Emanuele Paci[1,2]**

[1] School of Molecular and Cellular Biology, University of Leeds, Leeds, United Kingdom, [2] Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, United Kingdom, [3] School of Chemistry, University of Leeds, Leeds, United Kingdom, [4] Hepatopancreatobiliary Transplant Unit, St. James's University Hospital, Leeds, United Kingdom

## Abstract

The evolution of disease or the progress of recovery of a patient is a complex process, which depends on many factors. A quantitative description of this process in real-time by a single, clinically measurable parameter (biomarker) would be helpful for early, informed and targeted treatment. Organ transplantation is an eminent case in which the evolution of the post-operative clinical condition is highly dependent on the individual case. The quality of management and monitoring of patients after kidney transplant often determines the long-term outcome of the graft. Using NMR spectra of blood samples, taken at different time points from just before to a week after surgery, we have shown that a biomarker can be found that quantitatively monitors the evolution of a clinical condition. We demonstrate that this is possible if the dynamics of the process is considered explicitly: the biomarker is defined and determined as an optimal reaction coordinate that provides a quantitatively accurate description of the stochastic recovery dynamics. The method, originally developed for the analysis of protein folding dynamics, is rigorous, robust and general, i.e., it can be applied in principle to analyze any type of biological dynamics. Such predictive biomarkers will promote improvement of long-term graft survival after renal transplantation, and have potentially unlimited applications as diagnostic tools.

**Citation:** Krivov SV, Fenton H, Goldsmith PJ, Prasad RK, Fisher J, et al. (2014) Optimal Reaction Coordinate as a Biomarker for the Dynamics of Recovery from Kidney Transplant. PLoS Comput Biol 10(6): e1003685. doi:10.1371/journal.pcbi.1003685

**Editor:** Edwin Wang, National Research Council of Canada, Canada

**Received** March 10, 2014; **Accepted** May 8, 2014; **Published** June 26, 2014

**Data Availability:** The authors confirm that all data underlying the findings are fully available without restriction. The data can be found in the Supporting Information file Data S1.

**Funding:** SVK was supported by EPSRC grant EP/E500080/1. HF was supported by EPSRC grants 2004/05-EP/548622/1, 2005/06-EP/C512855/1, 2006/07-EP/E501869/1, 2007/08-EP/F500033/1, 2008/09-EP/G500010/1 www.epsrc.ac.uk. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* Email: s.krivov@leeds.ac.uk

## Introduction

The responses of an individual to an infection, to pharmacological treatment or to surgery are examples of time-dependent stochastic processes characterized by complex dynamics. An increasing amount of time-resolved data is available reporting on the unique chemical fingerprints that specific cellular processes leave behind [1,2]. Metabolites, such as those found in blood or urine, contain in principle a comprehensive picture (referred to as the metabolome) of the evolution of a patients condition. While such a picture is very complex and generally not insightful, the time evolution of the metabolome of a patient contains crucial information. Conventionally, a biomarker is sought by comparing differences in the metabolic profiles between two states (e.g. healthy and pathological) using unsupervised methods (such as principal component analysis, PCA [3]) or supervised methods (e.g., orthogonal projections to latent structures, OPLS [4] and related [5]). However, if one is interested in time-related changes to the metabolic profile, which are relevant to the pathological state, for example in the monitoring of disease progression or defining surrogate end points, the problem becomes more complex. A number of other methods, sometimes borrowed from other disciplines, have been proposed for analysis of time-resolved metabolomic data [2]; they rely in general on previous knowledge, either of the identity of the relevant metabolites and/or the functional form of the time dependence of their concentrations. When the underlying biochemical mechanism is itself unknown, such methods are obviously not useful.

Disease dynamics, according to the systems biology point of view, is more accurately described as dynamics of highly entangled molecular networks, with disease being an emerging property of the networks [6]. Adopting this view, we seek a biomarker, which is a descriptor (function) of the networks states, rather than of a few molecules. We assume that disease dynamics is a Markov (memory-less) stochastic process, in which future behavior is completely specified (in a probabilistic sense), by the current state of an organism, e.g., the complex of genome, proteome, metabolome, epigenome, age, environment, and whatever additional information may be required (hereafter the "configuration space"). Illustrative and enlightening is a recent study [1] where a combination of genomic, transcriptomic, proteomic, metabolomic and autoantibody profiles from a single individual was followed for over a 14 month period. The analysis uncovered extensive dynamics changes in diverse molecular components and biological

## Author Summary

The evolution of disease or the progress of recovery of a patient is usually monitored by collecting physical parameters, which may be simply the body temperature for a common cold or properties of tissue samples for e.g., cancer. Most often clinical decisions are taken based on the current value or because of a sizable change of a relevant parameter. As more advanced diagnostic tools become available, and huge numbers of parameters can be collected at short, frequent time intervals, two related questions arise. The first is, which of the parameters provides relevant information on the progress of disease or recovery as opposed to noise? Is there more information that can be obtained from the history of the evolution of such parameters? Here we propose a novel approach that leads, for the specific case of recovery from kidney transplant, to a positive answer.

pathways across healthy and disease states. In the case where the dynamics is stochastic rather than deterministic, a single observed trajectory is not sufficient for a complete description. In principle, a Markov state model that gives a complete description of the process can be constructed by observing various realizations of the disease, and computing the transition probabilities between all states. The model can be used to predict the properties of interest, for example, the probability of a given outcome (e.g., full recovery) after a certain time given an initial state. Such a straightforward approach cannot be realized in practice. The amount of information necessary to define exactly the state of an organism is huge and difficult to identify; the statistics necessary to construct the Markov state model grows exponentially with the dimensionality of the configuration space. Moreover, to be useful for practitioners (e.g., for diagnostic purposes), the description of the disease dynamics needs to be simplified. This can be done by introducing one or a few variables (hereafter the "reaction coordinate") that describe the properties of interest. Ideally, such a simplified description should be as predictive as the Markov state model previously described, i.e., the probability of a particular clinical outcome calculated from the value of the variable should closely approximate that computed based on the full Markov state model. If such properties are satisfied we call the variable an "optimal reaction coordinate".

### Determination of the optimal reaction coordinate

Here we present a general framework that allows us to determine such an optimal coordinate or biomarker from longitudinal cohort studies directly, without constructing the Markov state model. The method was originally developed to describe complex dynamics of protein folding [7–9]. Briefly, a putative functional form of the reaction coordinate is assumed, for example, a linear combination of features (here the metabolome $^1$H NMR spectra) that could describe the process. The approach is invariant to the choice of the functional form and the set of observables, provided they contain all the essential information about the dynamics of the process. The coordinate is optimized (trained) on a sample of trajectories, i.e., realizations of a complex multidimensional dynamical process. This is achieved by choosing the coordinate (e.g., the coefficients of the linear combination) such that the cut based free energy profile associated with the coordinate is the highest [7,10]. Namely, given an ensemble of N trajectories $\vec{X}_j(t)$ $(j=1,N)$ and a reaction coordinate functional form $y=R(\vec{X},\vec{\alpha})$ an ensemble of reaction coordinate trajectories is constructed by projecting the multidimensional trajectories onto the reaction coordinate $y_j(t)=R(\vec{X}_j(t),\vec{\alpha})$ $(j=1,N)$. The optimal reaction coordinate is found by optimizing the parameters $\vec{\alpha}$ so that the cut-based free energy profile $F_C(y)$ [7,8] is maximal. $F_C(y)=-\ln Z_C(y)$, where the partition function $Z_C(y)$ equals half number of transitions (crossings) by the reaction coordinate time-series through point $y$; here and below we set $k_BT=1$. The CFEP ($F_C$) unlike the conventional histogram based free energy profile ($F_H$) is invariant to reaction coordinate rescaling, insensitive to statistical noise and capable of detecting subdiffusion. Together they determine the coordinate dependent diffusion coefficient $D(x)$ and thus completely specify diffusive dynamics [9]. One can maximize instead the generalized cut based free energy profile $F_{C,1}(y)=-\ln Z_{C,1}(y)$ where the partition function $Z_{C,1}(y)$ takes into account each transition through point $y$ with weight equal to the transition distance; for a Gaussian distribution of steps (i.e., diffusive dynamics) the two optimality criteria are equivalent [10]. If the reaction coordinate is a weighted sum of basis functions $y=\sum_k \alpha_k X_k$, as used here, the optimal values of the parameters ($\alpha_i$ that maximize $F_{C,1}$ can be found analytically [10]).

In supervised optimization a coordinate that accurately describes the dynamics of transition between two given end states (e.g., healthy and disease) is determined. Incidentally, the coordinate is the probability of full recovery, i.e., of ending up in the "healthy" state rather than the "disease" state starting from a current state. It is known as committor or folding probability in protein folding studies [10,11]. If the two end states are separated by the highest free energy barrier, the transition between them corresponds to the slowest relaxation mode, and an eigenvector, corresponding to the slowest mode is an optimal reaction coordinate [10,12]. This coordinate can be determined in an unsupervised way without explicit definition of end states.

## Results/Discussion

The method outlined above (details given in Materials and Methods) has been used to analyze the evolution of $^1$H NMR spectra of the erythrocyte extracts of blood from 18 patients undergoing kidney transplantation; for each patient up to nine samples were taken before surgery and daily up to one week after. The spectra were normalized to the total sum of the spectral intensities and then coarse-grained with a bin size of 0.32 ppm. The average intensity within each bin was logarithmically transformed as $I_k=\log(10^6 I_k+1)$. The reaction coordinate was taken as a linear sum of the transformed average bin intensities: $y=\sum_k \alpha_k I_k$. The method to determine the optimal reaction coordinate is robust: it was repeated with different transformation (e.g., $I_k=\sqrt{I_k}$) or without transformation, with different bin sizes, in a supervised way, all leading to virtually identical results. Note however, as discussed below, a significant decrease of the bin size, e.g., to 0.1 ppm, while leading to a slightly better description results in over-fitting.

Independently of the NMR data, patients have been divided in three classes based on a clinical assessment of the patients into "primary function" (PF), "delayed graft function" (DGF) and "acute rejection" (AR) with and without primary function. Primary function was defined as immediate recovery of renal function following surgery. Delayed graft function was defined as the need for dialysis in the first week following transplantation.

The spectra for a single patient for nine different time points (a "trajectory"), are shown in Fig. 1a. The trajectory for each patient projected onto the first principal component (Fig. 1b) shows individual variability but no separation between different classes of

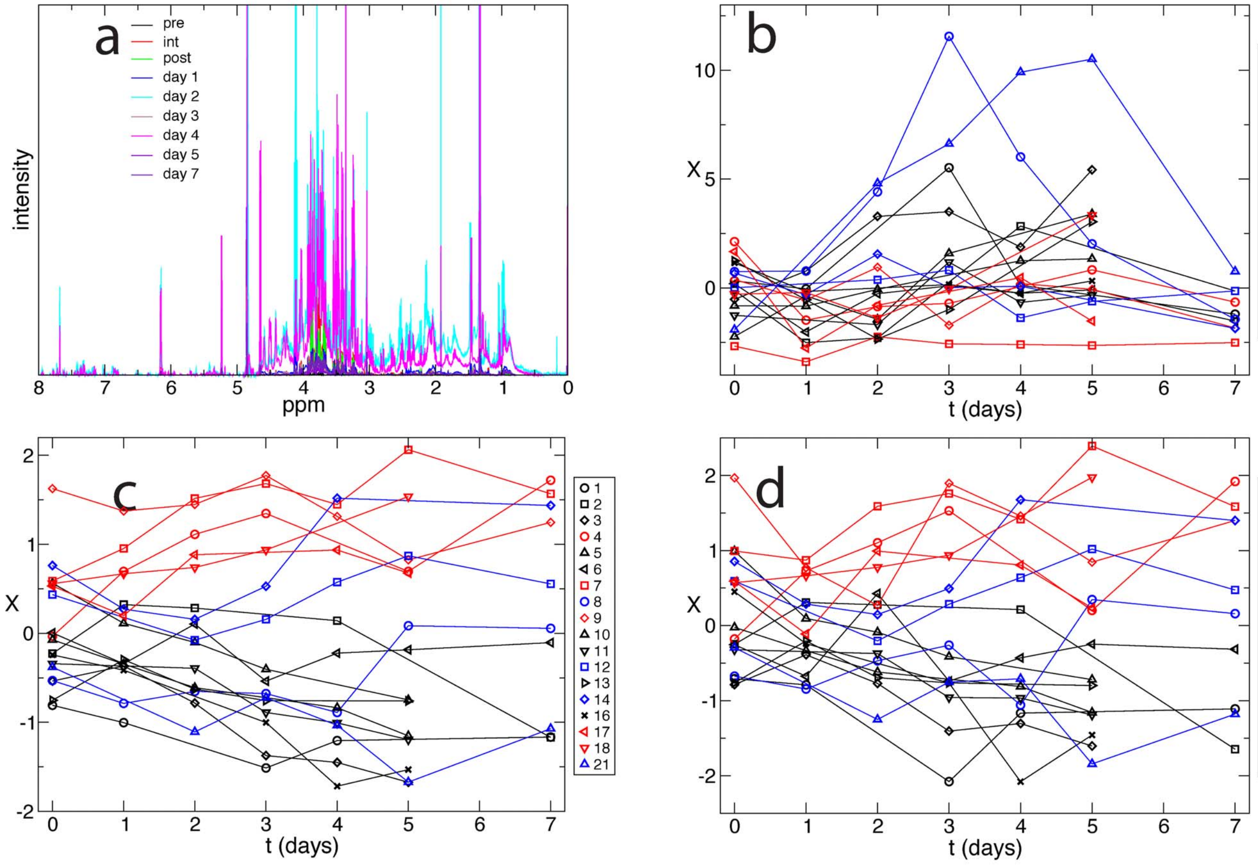

**Figure 1. Unsupervised optimization.** a) NMR spectra for blood extracts of a single patient, collected over nine time points. b) Patients trajectories projected on the first principal component do not show any feature enabling us to separate the trajectories according to the clinical assessment of the patients. The color indicates the final clinical classification of the patient: primary function patients are shown in black, delayed graft function in red, and acute rejection in blue. c) Patients trajectories projected on the optimal reaction coordinate. d) Leave-one-out cross-validation: every trajectory is projected on the optimal reaction coordinate constructed without that specific trajectory.
doi:10.1371/journal.pcbi.1003685.g001

patients, and no relevant information on the evolution of their clinical conditions during the period in which the samples were taken. The same is true if the trajectory is projected onto the first few principal components.

Fig. 1c shows patients trajectories projected onto the optimal reaction coordinate determined in the unsupervised way (the second eigenvector). Trajectories categorized as PF are characterized by an evolution towards negative values of the coordinate; those in the DGF cohort evolve towards positive values of the coordinate. The time evolution of the AR patients cannot be discerned within the timescale of the data collection. Importantly, the clinical group to which each patient could be ascribed to is apparent from about the second day after surgery, earlier than any other clinical indicator, including invasive, though gold-standard, biopsy. Note that separation between PF and DGF patients cannot be due to dialysis, since dialysis was never performed in the first two days after surgery.

The difference between the results of PCA (Fig. 1b) and of the approach proposed here (Fig. 1c) can be understood as follows. PCA (and other algorithms [10]) perform dimensionality reduction with a focus on representation of the properties of the configuration space. In particular, the PCA maximizes the variability of configurations along the principal components. The dynamical information contained in the temporal sequence of configurations (trajectory) is ignored. By explicitly considering the dynamics, our approach performs dimensionality reduction while preserving and exploiting the dynamics of the process [10,13].

To demonstrate that the results are not affected by over fitting (even though the analysis is unsupervised) the leave-one-out cross-validation procedure was performed. In Fig. 1d every trajectory is projected on the optimal reaction coordinate constructed without the trajectory. All the trajectories are in good agreement with those in Fig. 1c, and the prediction on the future evolution of each trajectory (i.e., the fate of each patient) is identical. This confirms that the constructed coordinate is robust and that the biomarker can be used to follow up the evolution of the condition of a new patient.

The leave-one-out cross-validation was also instrumental in the choice of the bin size. Decreasing the bin size increases the number of parameters and thus the flexibility of the coordinate; while this leads to a slightly better separation, the cross-validation test fails, a clear consequence of over fitting. Note that, in principle, the possibility of over fitting, as well as the optimality of the coordinate, could be established by comparing the cut profiles computed with different time intervals [8,14] or with other methods [15–17]. Alternatively, one may optimize with an over fitting penalty [8], eliminating the need in the manual choice of the bin size. Here we did not use such an approach, because the trajectories are too short to be sampled with larger intervals.

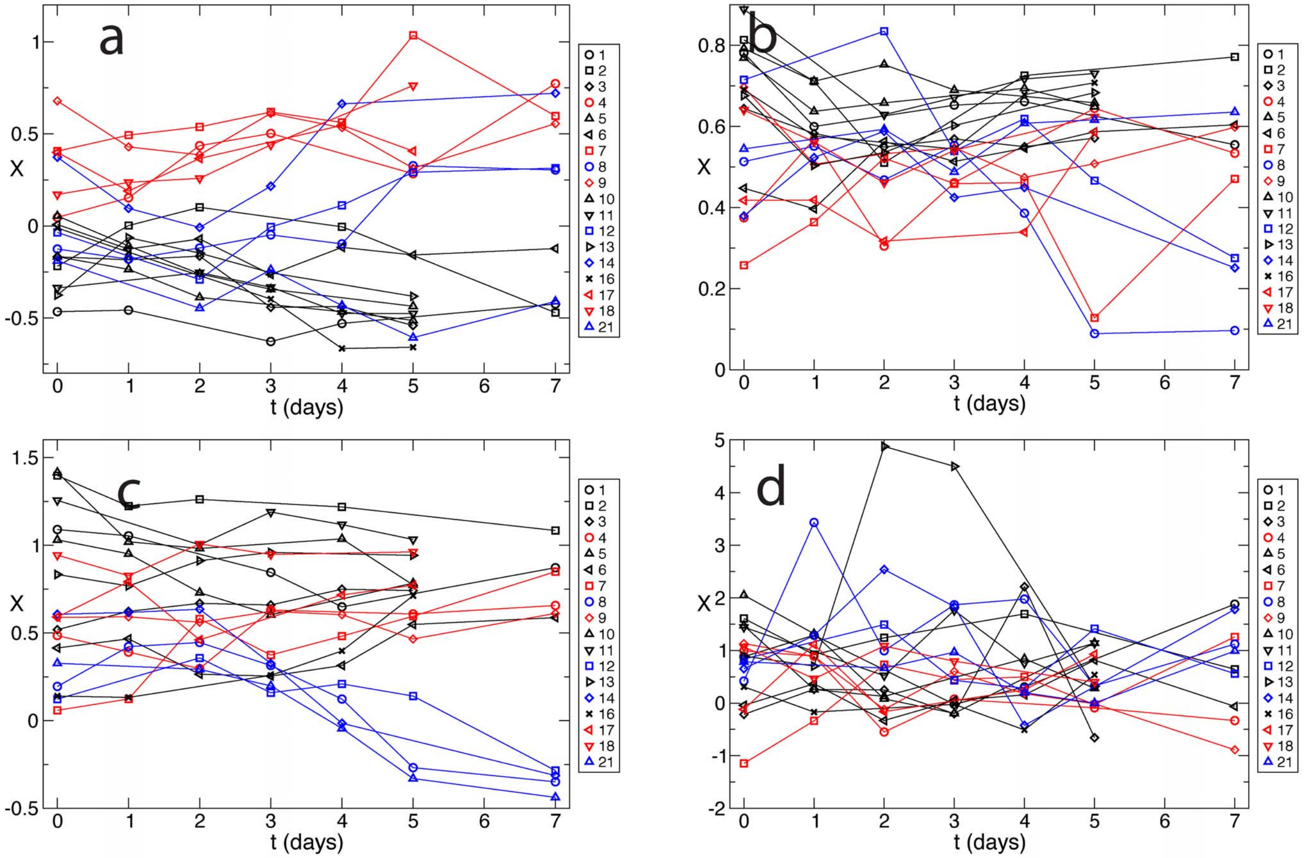

**Figure 2. Supervised optimisation. a)** PF versus AR and DGF with binning interval of 0.32 ppm: the results are similar to the unsupervised analysis Fig. 1C. **b)** AR versus PF and DGF with binning interval of 0.32 ppm: no clear separation between the classes. **c)** AR versus PF and DGF with binning interval of 0.1 ppm: AR is separated from PF and DGF. **d)** leave one out cross validation of panel **c**: small binning interval of 0.1 ppm leads to overfitting.
doi:10.1371/journal.pcbi.1003685.g002

While the second eigenvector describes the transitions dynamics over the highest barrier, it is possible that the third eigenvector, which describes the second slowest process, would separate AR cases from PF and/or DGF cases. However, this is not the case. Among the many possible reasons why AR cases cannot be discriminated from the other cases, one could be the insufficient flexibility of the reaction coordinate due to the small number of parameters. Increasing the number of parameters while avoiding over fitting would require more trajectories (i.e., more patients) than those available. To illustrate this fact we performed a supervised analysis aimed at separating AR from PF and/or DGF trajectories. Fig. 2a shows supervised optimization where PF trajectories were terminated at $-1$ and AR and DGF trajectories at 1. The results are virtually identical to those obtained by unsupervised optimization reported in Fig. 1C. An attempt to separate AR from PF and DGF with 0.32 ppm binning interval did not produce meaningful results (Fig. 2b). Decrease of the size of the binning interval to 0.1 ppm resulted in a more flexible reaction coordinate, so that the supervised separation of AR from PF and DGF become possible Fig. 2c. However, results of leave-one-out cross-validation, shown in Fig. 2d illustrate that such small binning interval leads to over fitting.

Having determined an optimal coordinate, the dynamics of disease as a whole can be described as diffusion on the free energy profile (Fig. 3) along the optimal coordinate. The latter has been rescaled so that the diffusion coefficient equals unity. Two basins (attractors) naturally emerge, one identifying the PF condition and one for DGF and AR conditions together. Starting from the top of the profile, a patients trajectory can fall either on the left (PF) or

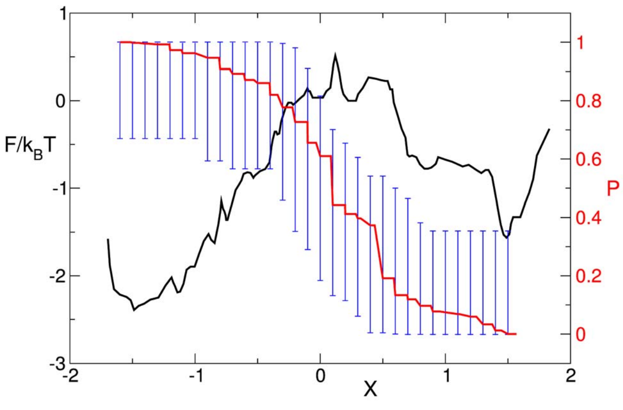

**Figure 3. The disease dynamics is described as diffusion on the free energy landscape (black).** The left and right basins correspond to PF and DGF+AR states, respectively. Probability of the successful outcome from current conditions P(x) computed from diffusion on the free energy landscape (red line) and directly from the trajectories (the blue vertical bar represents 95% confidence interval on the estimation).
doi:10.1371/journal.pcbi.1003685.g003

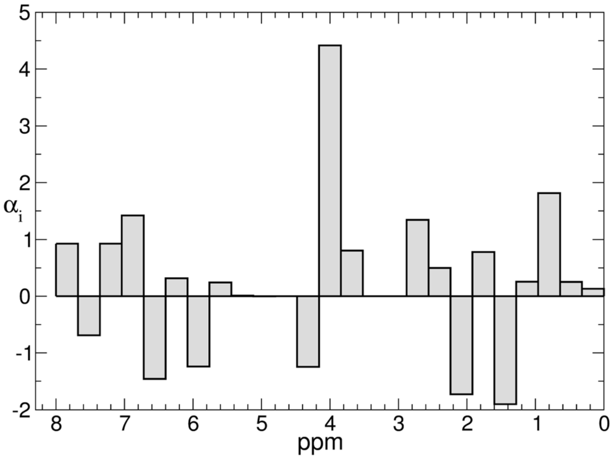

**Figure 4. Linear coefficients of the optimal reaction coordinate $y=\sum_k \alpha_k I_k$, where $I_k$ is the logarithm of the intensity of NMR spectra in bin k.**
doi:10.1371/journal.pcbi.1003685.g004

the right (DGF+AR) basin. Fig. 3 shows the probability of full recovery, i.e., of reaching the left basin (PF) before reaching the right one (DGF+AR) starting from any value of the coordinate (any condition before or after surgery) computed by assuming diffusive dynamics on the free energy profile and directly from the trajectories. Their agreement in combination with the cross-validation (Fig. 1d) demonstrates the possibility to predict the clinical outcome for patients not previously considered, meaning that the optimal coordinate found is a good biomarker.

The linear coefficients of the optimal reaction coordinate (Fig. 4) illustrate that the whole spectrum (metabolome) is important to determine the optimal coordinate that provides a fine-grained quantitatively accurate description of the dynamics. Indeed, the largest weight is associated with the bin ranging from 3.84 to 4.16 ppm; this includes signals for creatinine. Creatinine is currently used as a clinical marker in the assessment of renal function and to monitor post-transplant recovery [18]; the trend in the serum creatinine concentrations is more informative than absolute values, but it is by no means a specific marker. Other signals, e.g., at ca. 1 ppm/2 ppm and 1.4 ppm correspond to lipid and lactate respectively; the signal for the CH of lactate is at 4.2 ppm. We note that while the identity of each of the metabolites is undoubtedly interesting *per se*, their identities are not required for the usage of the optimal coordinate for diagnostic purposes as a biomarker.

In order to describe the specificity and sensitivity of a proposed biomarker one commonly uses the ROC analysis, which describes the trade-off between false positives and false negatives [19]. The proposed approach differs from the conventional ones (e.g., the PCA) by assuming that the disease dynamics is stochastic, which makes an application of such analysis not straightforward. This can be understood by considering a set of patients and a vector of measurable characteristics (e.g., the NMR spectra) for $i$-th patient denoted as $\mathbf{X}_i$. Conventionally, one assumes that the true state of a patient, denoted by $T_i$, can be unambiguously mapped (by a "gold standard" test) onto two classes $T_i=\{0,1\}$, e.g., healthy and diseased. And that this mapping is deterministic, i.e., that if $\mathbf{X}_i=\mathbf{X}_j$, then $T_i=T_j$. A biomarker is then a binary function $M(\mathbf{X})$ from $\mathbf{X}_i$ onto $\{0,1\}$ which best approximates $T_i$. Cases where $M(\mathbf{X}_i)\neq T_i$ correspond to the two possible types of errors - false positive and false negative. The former, for example, corresponds to the fraction of cases where $M(\mathbf{X}_i)=1$, while $T_i=0$. Since $T_i$ are determined by a "gold standard" test, the biomarker function is assumed to be the sole source of errors.

The proposed approach assumes that the current state of a patient $T_i$ is related to the two terminal states $\{0,1\}$, where the patient will end up eventually, only probabilistically. Identical patients with identical conditions undergoing identical treatment (i.e., $\mathbf{X}_i=\mathbf{X}$) will end up in different terminal states 1 and 0 with probabilities of $p$ and $1-p$, respectively. Thus knowing the terminal state does not allow one to determine the true value of the current state $T_i$, i.e., a "gold standard" test and stochastic dynamics are incompatible. Correspondingly, the purpose of a biomarker is not to approximate the terminal states but rather to approximate the probability $p(\mathbf{X}_i)$ of ending up in one of the two end states.

A way to asses the accuracy of such a biomarker is to judge how accurately it reproduces that probability, e.g., Fig. 3. The classification process is analogous to the Bernoulli trial of a binary random variable that accepts 1 with probability $p_i=p(\mathbf{X}_i)$. To determine the probability $p(\mathbf{X}_i)$ of ending up in the terminal state 1, starting from $\mathbf{X}_i$, one needs to repeat "the experiment" a number of times starting from the same conditions $\mathbf{X}_i$ (in particular, the same patient) and count the fraction of events ending up in 1. Such a direct approach is clearly unrealistic. An alternative is to combine the states $\mathbf{X}_i$ with similar $p(\mathbf{X}_i)\approx p$ and determine $p$ from such an ensemble of states, as have been done in Fig. 3.

## Conclusion

We have shown that the dynamics of recovery from kidney transplant can be quantitatively described as diffusion on a free energy profile, which is a function of a measurable biomarker. Such a biomarker can be determined in an (un)supervised way from longitudinal cohort studies (patient trajectories), which is optimal in the sense that it is able to discriminate where each patient is on a free energy profile. In particular, the probability of rapid recovery (primary function) can be used to devise optimal treatment. Such an approach is general and can be useful to develop optimal biomarkers for diseases that develop slowly and in a complicated way depending on many factors, or unknown unknowns, such as aging [20], cancer [21,22] and psychological disorders [23].

# Materials and Methods

## Ethics statement

Approval was given by the regional ethics approval committee approval number REC Ref: 07/H1306/129

## The cut-based free energy profiles

The partition function of the cut-based free energy profile $Z_{C,r}$ at point $y$ equals half the sum of the distances of those trajectory steps that go through point $y$ [14]. More precisely,

$$Z_{C,r}(y)=$$
$$1/2\sum_i |x(i\Delta t+\Delta t)-x(i\Delta t)|^r\Theta[-(x(i\Delta t+\Delta t)-y)(x(i\Delta t)-y)],$$

where $\Theta(x)$ is the Heaviside step function and $x(i\Delta t)$ is the reaction coordinate time series sampled with time interval $\Delta t$. The cut free energy profile is defined as $F_{C,r}(y)/kT=-\ln Z_{C,r}(y)$ and $F_C=F_{C,0}$; here we assume that $kT=1$.

The optimal coordinate is defined as the one with the highest cut profiles (lowest partition function). The justification of the optimization criteria can be summarized as follows (for the details see the cited references). It can be shown that minimum of $\int Z_{C,1}(y)dy$, with constrains $y(A)=0$ and $y(B)=1$, is attained when the reaction coordinate $y$ equals the $p_{fold}$ coordinate - an optimal coordinate [10]. Correspondingly, a sub-optimal coordinate with a lower value of the cut profile, has the mean square displacement which grows slower then linear with time [14]. The latter is an indication that dynamics is not diffusive and that non-Markovain memory effects are at play. Another manifestation of a sub-optimal coordinate, is that it has lower free energy barriers and thus a faster kinetics. The kinetics along the coordinate with the highest cut profile is the slowest [9,14].

### Supervised optimization

Optimization of the reaction coordinate can be performed in a supervised or unsupervised manner. In supervised optimization a coordinate that accurately describes the dynamics of transition between two given end states (e.g., healthy and disease) is determined. Incidentally, the optimal coordinate is the probability of full recovery, i.e., of ending up in the "healthy" state rather than the "disease" state starting from a current state. The optimization is constrained by fixing the value for the coordinate for the two state $y(A)=0$ and $y(B)=1$ [10]. If the reaction coordinate is a weighted sum of basis functions $y=R(\vec{X},\vec{\alpha})=\sum_k \alpha_k r_k(\vec{X})$, boundary conditions are given as $y_{i_0}=\sum_k \alpha_k r_k(\vec{X}(i_0\Delta t))=0$ and $y_{i_1}=\sum_k \alpha_k r_k(\vec{X}(i_1\Delta t))=1$ where $i_0$ and $i_1$ index the points which belong to A and B states, respectively. The optimal weights ($\alpha_i$) which give constrained maximum to $\int Z_{C,1}(y)dy$ can be found analytically [10]. Here we specified the constrains in the following way, which resulted in a more flexible coordinate. Instead of assuming that each trajectory ends at either 0 or 1, we assumed that each trajectory is constrained to end with either 0 or 1 (in other words a trajectory reaches an end state on the following day). In this case the optimal weights are found by minimizing $\int Z_{C,1}(y)dy \sim \sum_{i,j}[y_j(i\Delta t+\Delta t)-y_j(i\Delta t)]^2$, which equals

$$\sum_{j=1,N}\sum_{i=1,t_j-1}\left[\sum_k \alpha_k r_k(\vec{X}_j(i\Delta t+\Delta t))-\alpha_k r_k(\vec{X}_j(i\Delta t))\right]^2 + \sum_{j=1,N}\left[b_j-\sum_k \alpha_k r_k(\vec{X}_j(t_j))\right]^2,$$

where $i,j,k$ are indexes that refer to time frame, trajectory and basis function, respectively; the second term of the functional describes the boundary condition with $b_j$ equal 0 or 1 for trajectories connected to 0 or 1, respectively. The optimal parameters are found by solving the corresponding system of linear equations $\partial/\partial\alpha_k \int Z_{C,1}(y)dy=0$. To facilitate the visual comparison with the unsupervised results $b_j=0$ where changed to $b_j=-1$, which results in a shift and change of scale of the optimal coordinate.

### Unsupervised optimization

In unsupervised optimization the determined coordinate describes the slowest relaxation mode (the second eigenvector) of the stochastic dynamics [10]. If the two states ($A$ and $B$) are separated by the highest barrier, so the slowest relaxation rate corresponds to the transition dynamics between the states, the second eigenvector reaction coordinate approximates the folding probability (the probability of full recovery here) reaction coordinate in the transition state region - the most important part for the description of the transition dynamics [10,12]. The eigenvectors can be found by minimizing $I=\int Z_{C,1}(y)dy \sim \sum_{i,j}[y_j(i\Delta t+\Delta t)-y_j(i\Delta t)]^2$ under constraint $\sum_{i,j} y_j^2(i\Delta t)=1$ [10]. Due to the constraint, the optimization function simplifies to the auto-correlation function. If reaction coordinate is a weighted sum of basis functions, the optimal weights can be found analytically. They are the solution of the generalized eigenvalue problem [10].

### Determination of the equilibrium free energy profile and the probability of successful outcome

The free energy profile that describes the disease dynamics cannot be determined from the patients trajectory simply by computing the cut based (or histogram) free energy profiles because the trajectories are not at equilibrium. The procedure described in Ref [13] was employed. Briefly, assuming diffusive dynamics, the equilibrium free energy profile can be computed from the steady state (non-equilibrium) probability distribution $P_{ss}$ as

$$F(x)=-\ln P_{ss}(x)-\int^x \frac{J(x)dx}{D(x)P_{ss}(x)}.$$

Using $P_{ss}(x)=Z_H(x)$, $J(x)\Delta t=Z_C^+(x)-Z_C^-(x)$ and $D(x)=(Z_C/Z_H)^2\pi/\Delta t$, one obtains

$$F(x)=-\ln Z_H(x)-\int^x \frac{(Z_C^+(x)-Z_C^-(x))Z_H(x)dx}{\pi Z_C^2(x)},$$

where $Z_C^+(x)$ and $Z_C^-(x)$ are the cut profiles that measure flux in positive and negative direction, respectively.

Note that the method for determining the optimal reaction coordinate was originally derived for equilibrium dynamics; an extension of the framework to non-equilibrium dynamics has been suggested recently [24]. Here we assume that while non-equilibrium sampling affects populations, its main effect on the optimization procedure is in altering the contribution (weight) of the different regions to the optimization functional $\int Z_{C,1}(x)dx$ and can be neglected.

The probability of "full recovery" (the folding probability) was computed from the free energy profile as

$$p(x)=\frac{\int_x^b e^{F(x)}/D(x)dx}{\int_a^b e^{F(x)}/D(x)dx}$$

The success probability with 95% confidence interval were estimated from the trajectories by "add two successes and two failures" approach [25] as $\hat{p}_i \pm 2\sqrt{\hat{p}_i(1-\hat{p}_i)/(n_i^l+n_i^r+4)}$, where $\hat{p}_i=(n_i^l+2)/(n_i^l+n_i^r+4)$ and $n_i^l$ and $n_i^r$ are the numbers of trajectories visiting bin $i$ ended up in left or right half of the profile, respectively [26].

### Acquisition of NMR spectra

$^1$H NMR spectra were obtained for the water soluble components [27] of erythrocytes taken from 18 kidney transplant patients (up to 9 time points from pre-op to 7 days after surgery).

One-dimensional $^1$H NMR spectra were measured at 499.97 MHz on a Varian Unity Inova 500 spectrometer at 20°C, using a standard PRESAT pulse sequence. For all samples a relaxation delay of ca. 9 s (three times the longest T1) was applied between scans to allow the spins to fully relax, with 256 transients collected into 16384 data points and a spectral width of 6000 Hz.

An exponential line broadening of 0.5 Hz was applied to each free induction decay (FID) and zero filling to 32768 points was carried out, followed by Fourier transformation. Phase and baseline corrections were carried out using ACD/Labs 12.0 (Advanced Chemistry Development Inc., Toronto, Canada) and chemical shifts were referenced to the lactate doublet at 1.33 ppm.

### Clinical assessment of patients

Independently of the NMR data, patients have been divided in three classes based on a clinical assessment of the patients into primary function (PF), delayed graft function (DGF) and acute rejection (AR) with and without primary function. Primary function was defined as immediate recovery of renal function. Delayed graft function was defined as the need for dialysis in the first week following transplantation. Diagnosis of acute rejection was conducted on the basis of biopsy and histological findings. Dialysis was performed on day 5 to patient 4, day 2 to patient 7, day 4 to patient 9, day 7 to patient 17 and day 6 to patient 18. All biopsies were conducted between 6 and 9 days following transplantation. Nine patients had immediate primary function, five patients had delayed graft function and four patients had acute rejection. All but one transplant were eventually successful: in addition to acute rejection, patient 14 also suffered from renal artery stenosis, and the graft was ultimately removed. The immunosuppressive regime and induction agents were the same across the cohort.

## Supporting Information

**Data S1.** NMR Spectra of patients.
(TGZ)

## Author Contributions

Conceived and designed the experiments: SVK JF. Performed the experiments: HF PJG RKP. Analyzed the data: SVK EP. Wrote the paper: SVK HF PJG RKP JF EP.

## References

1. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, et al. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148: 1293–307.
2. Smilde AK, Westerhuis JA, Hoefsloot HC, Bijlsma S, Rubingh CM, et al. (2010) Dynamic metabolomic data analysis: a tutorial review. Metabolomics 6: 3–17.
3. Trygg J, Holmes E, Lundstedt T (2007) Chemometrics in metabonomics. Journal of proteome research 6: 469–79.
4. Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics 16: 119–128.
5. Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, et al. (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics 20: 341–351.
6. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. Nature 461: 218–23.
7. Krivov SV (2010) Is protein folding sub-diffusive? PLOS Comp Biol 6: e1000921.
8. Krivov SV (2011) The free energy landscape analysis of protein (FIP35) folding dynamics. J Phys Chem B 115: 12315–24.
9. Krivov SV, Karplus M (2008) Diffusive reaction dynamics on invariant free energy profiles. Proc Natl Acad Sci USA 105: 13841–6.
10. Krivov SV (2011) Numerical construction of the p-fold (committor) reaction coordinate for a markov process. J Phys Chem B 115: 11382–8.
11. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1998) On the transition coordinate for protein folding. J Chem Phys 108: 334–350.
12. Berezhkovskii A, Szabo A (2004) Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. J Chem Phys 121: 9186–7.
13. Krivov SV (2011) Optimal dimensionality reduction of complex dynamics: the chess game as diffusion on a free-energy landscape. Phys Rev E 84: 011135.
14. Krivov SV (2013) On reaction coordinate optimality. J Chem Theor Comput 9: 135–146.
15. Hummer G (2003) From transition paths to transition states and rate coefficients. J Chem Phys 120: 516–523.
16. Chodera JD, Pande VS (2011) Splitting probabilities as a test of reaction coordinate choice in single-molecule experiments. Phys Rev Let 107: 098102.
17. Peters B, Bolhuis PG, Mullen RG, Shea JE (2013) Reaction coordinates, one-dimensional smoluchowski equations, and a test for dynamical self-consistency. J Chem Phys 138: 054106.
18. Stenlund H, Madsen R, Vivi A, Calderisi M, Lundstedt T, et al. (2009) Monitoring kidney-transplant patients using metabolomics and dynamic modeling. Chemometrics and Intelligent Laboratory Systems 98: 45–50.
19. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical chemistry 4: 561–577.
20. Vaupel JW (2010) Biodemography of human ageing. Nature 464: 536–42.
21. Haeno H, Gonen M, Davis MB, Herman JM, Iacobuzio-Donahue CA, et al. (2012) Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies. Cell 148: 362–75.
22. Sawyers CL (2008) The cancer biomarker problem. Nature 452: 548–52.
23. Singh I, Rose N (2009) Biomarkers in psychiatry. Nature 460: 202–7.
24. Krivov SV (2013) Method to describe stochastic dynamics using an optimal coordinate. Phys Rev E 88: 062131.
25. Agresti A, Caffo B (2000) Simple and effective confidence intervals for proportions and dierences of proportions result from adding two successes and two failures. The American Statistician 54: 280–288.
26. Rao F, Settanni G, Guarnera E, Caisch A (2005) Estimation of protein folding probability from equilibrium simulations. J Chem Phys 122: 184901.
27. Cohn JN (2004) Introduction to surrogate markers. Circulation 109: IV20–1.