Bacterial Community Reconstruction Using Compressed Sensing

Amnon Amir[1,⋆] and Or Zuk[2,⋆]

[1] Department of Physics of Complex Systems,
Weizmann Institute of Science, Rehovot, Israel
[2] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
amnon.amir@weizmann.ac.il, orzuk@broadinstitute.org

Abstract. Bacteria are the unseen majority on our planet, with millions of species and comprising most of the living protoplasm. We propose a novel approach for reconstruction of the composition of an unknown mixture of bacteria using a single Sanger-sequencing reaction of the mixture. Our method is based on compressive sensing theory, which deals with reconstruction of a sparse signal using a small number of measurements. Utilizing the fact that in many cases each bacterial community is comprised of a small subset of all known bacterial species, we show the feasibility of this approach for determining the composition of a bacterial mixture. Using simulations, we show that sequencing a few hundred base-pairs of the 16S rRNA gene sequence may provide enough information for reconstruction of mixtures containing tens of species, out of tens of thousands, even in the presence of realistic measurement noise. Finally, we show initial promising results when applying our method for the reconstruction of a toy experimental mixture with five species. Our approach may have a potential for a simple and efficient way for identifying bacterial species compositions in biological samples.

Availability: supplementary information, data and MATLAB code are available at: http://www.broadinstitute.org/~orzuk/publications/BCS/

1 Introduction

Microorganisms are present almost everywhere on earth. The population of bacteria found in most natural environments consists of multiple species, mutually affecting each other, and creating complex ecological systems [28]. In the human body, the number of bacterial cells is over an order of magnitude larger than the number of human cells [37], with typically several hundred species identified in a given sample taken from humans (for example, over 400 species were characterized in the human gut [17], while [38] estimates a higher number of 500-1000, and 500 to 600 species were found in the oral cavity [36, 13]). Changes in the human bacterial community composition are associated with physical condition,

⋆ These authors contributed equally to this work.

V. Bafna and S.C. Sahinalp (Eds.): RECOMB 2011, LNBI 6577, pp. 1–15, 2011.

and may indicate [33] as well as cause or prevent various microbial diseases [22]. In a broader aspect, studies of bacterial communities range from understanding the plant-microbe interactions [40], to temporal and meteorological effects on the composition of urban aerosols [4], and is a highly active field of research [35].

Identification of the bacteria present in a given sample is not a simple task, and technical limitations impede large scale quantitative surveys of bacterial community compositions. Since the vast majority of bacterial species are non-amenable to standard laboratory cultivation procedures [1], much attention has been given to culture-independent methods. The golden standard of microbial population analysis has been direct Sanger sequencing of the ribosomal 16S subunit gene (16S rRNA) [25]. However, the sensitivity of this method is determined by the number of sequencing reactions, and therefore requires hundreds of sequences for each sample analyzed. A modification of this method for identification of small mixtures of bacteria using a single Sanger sequence has been suggested [29] and showed promising results when reconstructing mixtures of 2-3 bacteria from a given database of $\sim$260 human pathogen sequences.

Recently, DNA microarray-based methods [21] and identification via next generation sequencing (reviewed in [23]) have been used for bacterial community reconstruction. In microarray based methods, such as the Affymetrix PhyloChip platform [4], the sample 16S rRNA is hybridized with short probes aimed at identification of known microbes at various taxonomy levels. While being more sensitive and cheaper than standard cloning and sequencing techniques, each bacterial mixture sample still needs to be hybridized against a microarray, thus the cost of such methods limit their use for wide scale studies. Methods based on next generation sequencing obtain a very large number of reads of a short hyper-variable region of the 16S rRNA gene [2, 12, 24]. Usage of such methods, combined with DNA barcoding, enables high throughput identification of bacterial communities, and can potentially detect species present at very low frequencies. However, since such sequencing methods are limited to relatively short read lengths (typically a few dozens and at most a few hundred bases in each sequence), the identification is non unique and limited in resolution, with reliable identification typically up to the genus level [26]. Improving resolution depends on obtaining longer read lengths, which is currently technologically challenging, and/or developing novel analytical methods which utilize the (possibly limited) information from each read to allow in aggregate a better separation between the species.

In this work we suggest a novel experimental and computational approach for sequencing-based profiling of bacterial communities (see Figure 1). We demonstrate our method using a single Sanger sequencing reaction for a bacterial mixture, which results in a linear combination of the constituent sequences. Using this mixed chromatogram as linear constraints, the sequences which constitute the original mixture are selected using a Compressed Sensing (**CS**) framework.

Compressed Sensing (**CS**) [5, 14] is an emerging field of research, based on statistics and optimization with a wide variety of applications. The goal of **CS** is recovery of a signal from a small number of measurements, by exploiting the fact that many natural signals are in fact sparse when represented at a certain

appropriate basis. Compressed Sensing designs sampling techniques that condense the information of a compressible signal into a small amount of data. This offers the possibility of performing fewer measurements than were thought to be needed before, thus lowering costs and simplifying data-acquisition methods for various types of signals in many distantly related fields such as magnetic resonance imaging [32], single pixel camera [16], geophysics [30] and astronomy [3]. Recently, **CS** has been applied to various problems in computational biology, e.g. for pooling designs for re-sequencing experiments [18,39], for drug-screenings [27] and for designing multiplexed DNA microarrays [10], where each spot is a combination of several different probes.

The classical **CS** problem is solving the under-determined linear system,

$$\mathcal{A}\mathbf{v} = \mathbf{b} \tag{1}$$

where $\mathbf{v} = (v_1, ..., v_N)$ is the vector of unknown variables, $\mathcal{A}$ is the *sensing* matrix, often called also the *mixing* matrix and $\mathbf{b} = (b_1, ..., b_k)$ are the measured values of the k equations. The number of variables N, is far greater than the number of equations k. Without further information, $\mathbf{v}$ cannot be reconstructed uniquely since the system is under-determined. Here one uses an additional sparsity assumption on the solution - by assuming that we are interested only in solution vectors $\mathbf{v}$ with only at most s non-zero entries, for some $s \ll N$. According to the **CS** theory, when the matrix $\mathcal{A}$ satisfy certain conditions, most notably the Restricted Isometry Property(RIP) [6,7], one can find the sparsest solution uniquely by using only a logarithmic number of equations, $k = O(s \log(N/s))$, instead of a linear number (N) needed for general solution of a linear system. Briefly, RIP for a matrix $\mathcal{A}$ means that any subset of $2s$ columns of $\mathcal{A}$ is almost orthogonal (although since $k < N$ the columns cannot be perfectly orthogonal). This property makes the matrix $\mathcal{A}$ invertible for sparse vectors v with sparsity s, and allows accurate recovery of $\mathbf{v}$ from eq. (1) - for more details on the RIP condition and the reconstruction guarantees see [6,7].

In this paper, we show an efficient application of pooled Sanger-sequencing for bacterial communities reconstruction using **CS**. The sparsity assumption is fulfilled by noting that although numerous species of bacteria have been characterized and are present on earth, at a given sample typically only a small fraction of them are present at significant levels. The proposed Bacterial Compressed Sensing (**BCS**) algorithm uses as inputs a database of known 16S rRNA sequences and a single Sanger-sequence of the unknown mixture, and returns the sparse set of bacteria present in the mixture and their predicted frequencies. We show a successful reconstruction of simulated mixtures containing dozens of bacterial species out of a database of tens of thousands, using realistic biological parameters. In addition, we demonstrate the applicability of our method for a real sequencing experiment using a toy mixture of five bacterial species.

2 The BCS Algorithm

In the Bacterial Community Reconstruction Problem we are given a bacterial mixture of unknown composition. In addition, we have at hand a database of the

orthologous genomic sequences for a specific known gene, which is assumed to be present in a large number of bacterial species (in our case, the gene used was the 16S rRNA gene). Our purpose is to reconstruct the identity of species present in the mixture, as well as their frequencies, where the assumption is that the sequences for the gene in all or the vast majority of species present in the mixture are available in the database. The input to the reconstruction algorithm is the measured Sanger sequence of the gene in the mixture (see Figure 1). Since Sanger sequencing proceeds independently for each DNA strand present in the sample, the sequence chromatogram of the mixture corresponds to the linear combination of the constituent sequences, where the linear coefficients are proportional to the abundance of each species in the mixture.

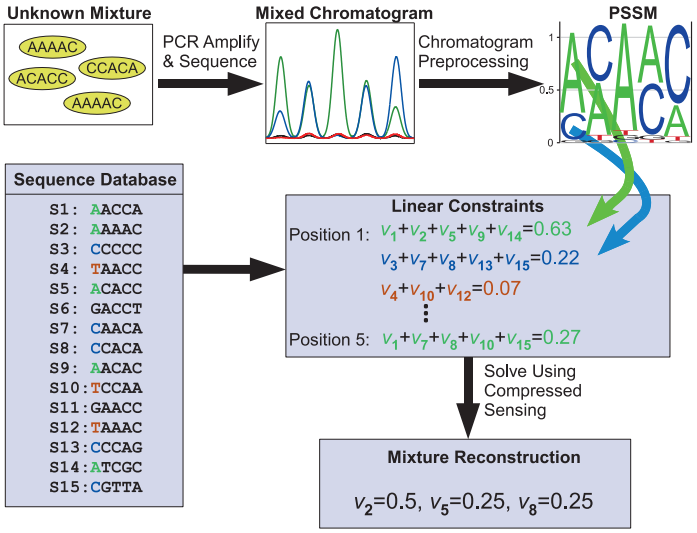

Fig. 1. Schematics of the proposed BCS reconstruction method. The 16S rRNA gene is PCR-amplified from the mixture and then subjected to Sanger sequencing. The resulting chromatogram is preprocessed to create the Position Specific Score Matrix (PSSM). For each sequence position, four linear mixture equations are derived from the 16S rRNA sequence database, with v_i denoting the frequency of sequence i in the mixture, and the frequency sum taken from the experimental PSSM. These linear constraints are used as input to the **CS** algorithm, which returns the sparsest set of bacteria recreating the observed PSSM.

Let N be the number of known bacterial species present in our database. Each bacterial population is characterized by a vector $\mathbf{v} = (v_1, ..., v_N)$ of frequencies of the different species. Denote by $s = \|\mathbf{v}\|_{\ell 0}$ the number of species present in the sample, where $\|.\|_{\ell 0}$ is the ℓ_0 norm which simply counts the number of non-zero elements of a vector $\|\mathbf{v}\|_{\ell 0} = \sum_i 1_{\{v_i \neq 0\}}$. While the total number of known species N is usually very large (in our case on the order of tens to hundreds of thousands), a typical bacterial community consists of a small subset of the

species, and therefore in a given sample, $s \ll N$, and $\mathbf{v}$ is a sparse vector. The database sequences are denoted by a matrix S, where S_{ij} is the j'th nucleotide in the orthologous sequence of the i'th species ($i = 1, .., N, j = 1, .., k$).

We represent the results of the mixture Sanger sequencing as a $4 \times k$ Position-specific-Score-Matrix (PSSM) comprised of the four vectors $\mathbf{a}, \mathbf{c}, \mathbf{g}, \mathbf{t}$, representing the measured frequencies of the four nucleotides in sequence positions $1..k$. The frequency of each nucleotide at a given position j gives a linear constraint on the mixture:

$$\sum_{i=1}^{N} v_i 1_{\{S_{ij}=`A'\}} = a_j \tag{2}$$

and similarly for the nucleotides 'C','G' and 'T'.

Define the $k \times N$ mixture matrix A for the nucleotide 'A':

$$A_{ij} = \begin{cases} 1 \; S_{ij} = `A' \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

and similarly for the nucleotides 'C','G','T'. The constraints given by the sequencing reaction can therefore be expressed in matrix form as:

$$A\mathbf{v} = \mathbf{a}, C\mathbf{v} = \mathbf{c}, G\mathbf{v} = \mathbf{g}, T\mathbf{v} = \mathbf{t} \tag{4}$$

The crucial assumption we make in order to cope with the insufficiency of information is the sparsity of the vector $\mathbf{v}$, which reflects the fact that only a small number of species are present in the mixture. We therefore seek a sparse solution for the set of equations (4). **CS** theory shows that under certain conditions on the mixture matrix and the number of measurements (see below), the sparse solution can be recovered uniquely by solving the following minimization problem [8, 15, 41]:

$$\mathbf{v}^* = \underset{\mathbf{v}}{argmin} \|\mathbf{v}\|_{\ell 1} = \underset{\mathbf{v}}{argmin} \sum_{i=1}^{N} |v_i| \quad s.t. \quad A\mathbf{v} = \mathbf{a}, C\mathbf{v} = \mathbf{c}, G\mathbf{v} = \mathbf{g}, T\mathbf{v} = \mathbf{t} \tag{5}$$

which is a convex optimization problem whose solution can be obtained in polynomial time. The above formulation requires our measurements to be precisely equal to their expected value based on the species frequency and the linearity assumption for the measured chromatogram. This description ignores the effects of noise, which is typically encountered in practice, on the reconstruction. Clearly, measurements of the signal mixtures suffer from various types of noise and biases. Fortunately, the **CS** paradigm is known to be robust to measurement noise [6,9]. One can cope with noise by enabling a trade-off between sparsity and accuracy in the reconstruction merit function, which in our case is formulated as:

$$\mathbf{v}^* = \underset{\mathbf{v}}{argmin} \frac{1}{2}\left(\|\mathbf{a} - A\mathbf{v}\|_{\ell 2}^2 + \|\mathbf{c} - C\mathbf{v}\|_{\ell 2}^2 + \|\mathbf{g} - G\mathbf{v}\|_{\ell 2}^2 + \|\mathbf{t} - T\mathbf{v}\|_{\ell 2}^2\right) + \tau\|\mathbf{v}\|_{\ell 1} \tag{6}$$

This problem represents a more general form of eq. (5), and accounts for noise in the measurement process. This is utilized by insertion of an $\ell 2$ quadratic error term. The parameter τ determines the relative weight of the error term vs. the sparsity promoting term. Many algorithms which enable an efficient solution of problem (6) are available, and we have chosen the widely used GPSR algorithm described in [19]. The error tolerance parameter was set to $\tau = 10$ for the simulated mixture reconstruction, and $\tau = 100$ for the reconstruction of the experimental mixture. These values achieved a rather sparse solution in most cases (a few species reconstructed with frequencies above zero), while still giving a good sensitivity. The performance of the algorithm was quite robust to the specific value of τ used, and therefore further optimization of the results by fine tuning τ was not followed in this study.

3 Results

3.1 Simulation Results

In order to asses the performance of the proposed **BCS** reconstruction algorithm, random subsets of species from the greengene database [11] were selected. Within these subsets, the relative frequencies of each species were drawn at random from a uniform frequency distribution normalized to sum to one (results for a different, power-law frequency distribution, are shown later), and the mixture Sanger-sequence PSSM was calculated . This PSSM was then used as the input for the **BCS** algorithm, which returned the frequencies of database sequences predicted to participate in the mixture (see Figure 1 and online Supplementary Methods).

A sample of a random mixture of 10 sequences, and a part of the corresponding mixed sequence PSSM, are shown in Figure 2A,B respectively. Results of the **BCS** reconstruction using a 500 bp long sequence are shown in Figure 2C. The **BCS** algorithm successfully identified all of the species present in the original mixture, as well as several false positives (species not present in the original mixture). The largest false positive frequency was 0.01, with a total fraction of 0.04 false positives. In order to quantify the performance of the **BCS** algorithm, we used two main measures: RMSE and recall/precision. RMSE is the Root-Mean Squared-Error between the original mixture vector and the reconstructed vector, defined as $RMSE(\mathbf{v}, \mathbf{v}^*) = \|\mathbf{v} - \mathbf{v}^*\|_{\ell 2} = \left(\sum_{i=1}^{N} (v_i - v_i^*)^2 \right)^{1/2}$. This measure accounts both for the presence or absence of species in the mixture, as well as their frequencies. In the example shown in Figure 2 the RMSE score of the reconstruction was 0.03. As another measure, we have recorded the *recall*, defined as the fraction of species present in the original vector $\mathbf{v}$ which were also present in the reconstructed vector $\mathbf{v}^*$ (this is also known as sensitivity), and the *precision*, defined as the fraction of species present in the reconstructed vector $\mathbf{v}^*$ which were also present in the original mixture vector $\mathbf{v}$. Since the predicted frequency is a continuous variable, whereas the recall/precision relies on a binary categorization, a minimal threshold for calling a species present in the reconstructed mixture was used before calculating the recall/precision scores.

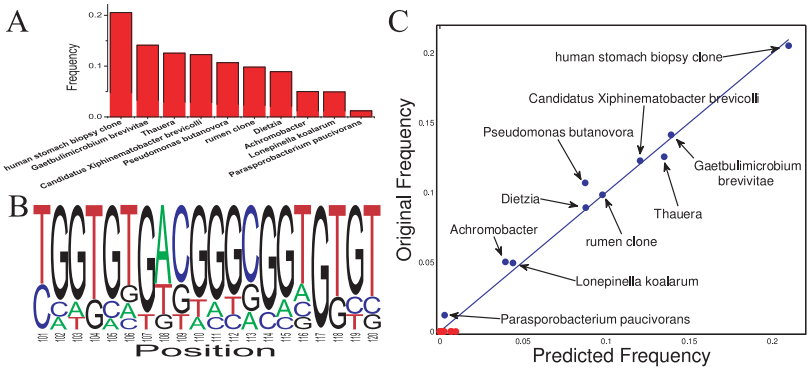

Fig. 2. Sample reconstruction of a simulated mixture. A. Frequencies and species for a simulated random mixture of $s = 10$ sequences. Species were randomly selected from the 16S rRNA database, with frequencies generated from a uniform distribution. **B.** A 20 nucleotide sample region of the PSSM for the mixture in (A). **C.** True vs. predicted frequencies for a sample **BCS** reconstruction for the mixture in (A) using $k = 500$ bases of the simulated mixture. Red circles denote species returned by the **BCS** algorithm which are not present in the original mixture.

Effect of Sequence Length. To determine the typical sequence length required for reconstruction, we tested the **BCS** algorithm performance using different sequence lengths. In Figure 3A (black line) we plot the reconstruction RMSE for random mixtures of 10 species. To enable faster running times, each simulation used a random subset of $N = 5000$ sequences from the sequence database for mixture generation and reconstruction. It is shown in Figure 3A that using longer sequence lengths results in a larger number of linear constraints and therefore higher accuracy, with $\sim$300 nucleotides sufficing for accurate reconstruction of a mixture of 10 sequences. The large standard deviation is due to a small probability of selection of a similar but incorrect sequence in the reconstruction, which leads to a high RMSE. Due to a cumulative drift in the chromatogram peak position prediction, typical usable experimental chromatogram lengths are in the order of $k \sim 500$ bases rather than the $\sim$1000 bases usually obtained when sequencing a single species (see online Supplementary Methods for details).

In order to asses the effect of similarites between the database sequences (which leads to high coherence of the mixing matrix columns) on the performance of the **BCS** algorithm, a similar mixture simulation was performed using a database of random nucleotide sequences (i.e. each sequence was composed of i.i.d. nucleotides with 0.25 probability for 'A','C','G' or 'T'). Using a mixing matrix derived from these random sequences, the **BCS** algorithm showed better performance (green line in Figure 3A), with $\sim$ 100 nucleotides sufficing for a similar RMSE as that obtained for the 16S rRNA database using 300 nucleotides.

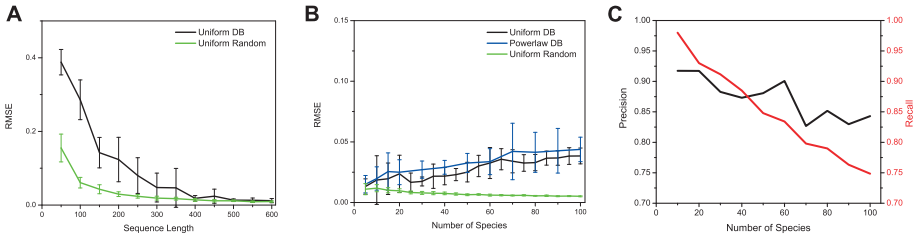

Fig. 3. Reconstruction of simulated mixtures. A. Effect of sequence length on reconstruction performance. RMSE between the original and reconstructed frequency vectors for uniformly distributed random mixtures of $s = 10$ species from the 16S rRNA database (black) or randomly generated sequences (green). Error bars denote the standard deviation derived from 20 simulations. **B.** Dependence of reconstruction performance on number of species in the mixture. Simulation is similar to (A) but using a fixed sequence length ($k = 500$) and varying the number of species in the mixture. Blue line shows reconstruction performance on a mixture with power-law distributed species frequencies ($v_i \sim i^{-1}$). **C.** Recall (fraction of sequences in the mixtures identified, shown in red) and precision (fraction of incorrect sequences identified, shown in black) of the **BCS** reconstruction of uniformly distributed database mixtures shown as black line in (B). The minimal reconstructed frequency for a species to be declared as present in the mixture was set to 0.25%.

Effect of Number of Species. For a fixed value of $k = 500$ nucleotides per sequencing run, the effect of the number of species present in the mixture on reconstruction performance is shown in Figure 3B,C. Even on a mixture of 100 species, the reconstruction showed an average RMSE less than 0.04, with the highest false positive reconstructed frequency (i.e. frequency for species not present in the original mixture) being less than 0.01. Using a minimal frequency threshold of 0.0025 for calling a species present in the reconstruction, the **BCS** algorithm shows an average recall of 0.75 and a precision of 0.85. Therefore, while the sequence database did not perform as well as random sequences, the 16S rRNA sequences exhibit enough variation to enable a successful reconstruction of mixtures of tens of species with a small percent of errors.

The frequencies of species in a biologically relevant mixture need not be uniformly distributed. For example, the frequency of species found on the human skin [20] were shown to resemble a power-law distribution. We therefore tested the performance of the **BCS** reconstruction on a similar power-law distribution of species frequencies with with $v_i \sim i^{-1}$. Performance on such a power-law mixture is similar to the uniformly distibuted mixture (blue and green lines in Figure 3B respectively) in terms of the RMSE. A sample power-law mixture and reconstruction are shown in Figure S4A,B. The recall/precision of the **BCS** algorithm on such mixtures (Figure S4C) is similar to the uniform distribution for mixtures containing up to 50 species, with degrading performance on larger mixtures, due to the long tail of low frequency species.

Effect of Noise on BCS Solution. Experimental Sanger sequencing chromatograms contain inherent noise, and we cannot expect to obtain exact measurements in practice. We therefore turned to study the effect of noise on the accuracy of the **BCS** reconstruction algorithm. Measurement noise was modeled as additive i.i.d. Gaussian noise $z_{ij} \sim N(0, \sigma^2)$ applied to each nucleotide read at each position. Noise is compensated for by the insertion of the $\ell 2$ norm into the minimization problem (see eq. (6)), where the factor τ determines the balance between sparsity and error-tolerance of the solution. The effect of added random i.i.d. Gaussian noise to each nucleotide measurement is shown in Figure 4. The reconstruction performance slowly degrades with added noise both for the real 16S rRNA and the random sequence database.

Using a noise standard deviation of $\sigma = 0.15$ (which is the approximate experimental noise level - see later) and sequencing 500 nucleotides, the reconstruction

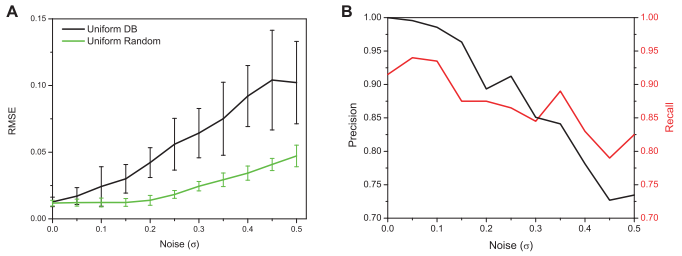

Fig. 4. Effect of noise on reconstruction. A. Reconstruction RMSE of mixtures of $s = 10$ sequences of length $k = 500$ from the 16S rRNA sequence database (black) or random sequences (green), with Gaussian noise added to the chromatogram. **B.** Recall (red) and precision (black) of the 16S rRNA database mixture reconstruction shown in (A).

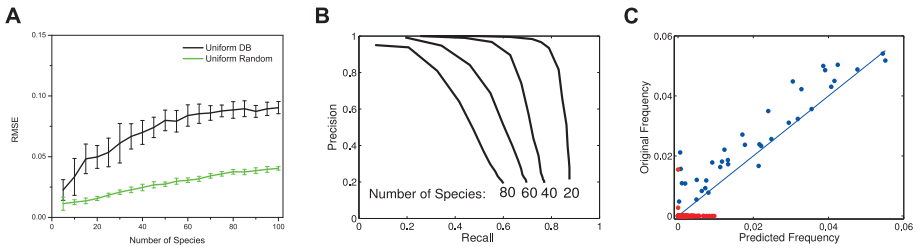

Fig. 5. Reconstruction with experimental noise level. A. Reconstruction RMSE as a function of number of species present in the mixture. Frequencies were sampled from a uniform distribution. Noise is set to $\sigma = 0.15$. Sequence length is set to $k = 500$. Black and green lines represent 16S rRNA and random sequences respectively. **B.** Recall vs. precision curves for different number of 16S rRNA sequences as in (A) obtained by varying the minimal inclusion frequency threshold. **C.** Sample reconstruction of $s = 40$ 16S rRNA sequences from (A).

performance as a function of the number of species in the mixture is shown in Figure 5. Under this noise level, the **BCS** algorithm reconstructed a mixture of 40 sequences with an average RMSE of 0.07 (Figure 5B), compared to ~ 0.02 when no noise is present (Figure 3B). By using a minimal frequency threshold of 0.006 for the predicted mixture, **BCS** showed a recall (sensitivity) of ~ 0.7, with a precision of ~ 0.7 (see Figure 5B), attained under realistic noise levels. To conclude, we have observed that the addition of noise leads to a graceful degradation in the reconstruction performance, and one can still achieve accurate reconstruction with realistic noise levels.

3.2 Reconstruction of an Experimental Mixture

While these simulations show promising results, they are based on correctly converting the experimentally measured chromatogram to the PSSM used as input to the **BCS** algorithm (see Figure 1). A major problem in this conversion is the large variability in the peak heights and positions observed in Sanger sequencing chromatograms (see Figure S2). It has been previously shown that a large part of this variability stems from local sequence effects on the polymerase activity [31]. In order to overcome this problem, we utilize the fact that both peak position and height are local sequence dependent, in order to accurately predict the chromatograms of the sequences present in the 16S rRNA database. The **CS** problem is then stated in terms of reconstruction of the measured chromatogram using a sparse subset of predicted chromatograms for the 16S rRNA database. This is achieved by binning both the predicted chromatograms and the measured mixture chromatogram into constant sized bins, and applying the **BCS** algorithm on these bins (see online Supplementary Methods and Figure S1).

We tested the feasibility of the **BCS** algorithm on experimental data by reconstructing a simple bacterial population using a single Sanger sequencing chromatogram. We used a mixture of five different bacteria: (*Escherichia coli W3110, Vibrio fischeri, Staphylococcus epidermidis, Enterococcus faecalis* and *Photobacterium leiognathi*). A sample of the measured chromatogram is shown in Figure 6A (solid lines). The **BCS** algorithm relies on accurate prediction of the chromatograms of each known database 16S rRNA sequence. In order to asses the accuracy of these predictions, Figure 6A shows a part of the predicted chromatogram of the mixture (dotted lines) which shows similar peak positions and heights to the ones experimentally measured (solid lines). The sequence position dependency of the prediction error is shown in Figure 6B. On the region of bins 125-700 the prediction shows high accuracy, with an average root square error of 0.08. The loss of accuracy at longer sequence positions stems from a cumulative drift in predicted peak positions, as well as reduced measurement accuracy. We therefore used the region of bins 125-700 for the **BCS** reconstruction.

Results of the reconstruction are shown in Figure 6C. The algorithm successfully identifies three of the five bacteria (*Vibrio fischeri, Enterococcus faecalis* and *Photobacterium leiognathi*). Out of the two remaining strains, one (*Staphylococcus epidermidis*) is identified at the genus level, and the other (*Escherichia coli*) is mistakenly identified as *Salmonella enterica*. While *Escherichia coli* and

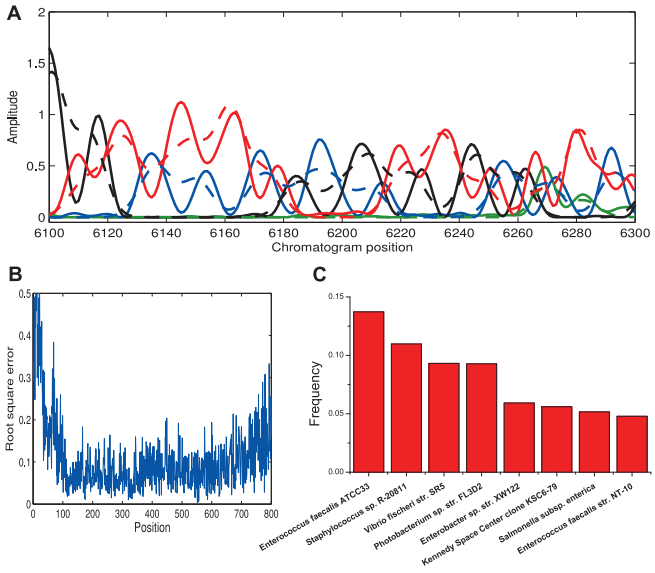

Fig. 6. Reconstruction of an experimental mixture. A. Sample region of the mixed chromatogram (solid lines). 16S rRNA from five bacteria was extracted and mixed at equal proportions. Dotted lines show the local-sequence corrected prediction of the chromatogram using the known mixture sequences. **B.** Square root distance between the predicted and measured chromatograms shown in (A) as a function of bin position, representing nucleotide position in the sequence. Prediction error was low for sequence positions $\sim 100-700$. **C.** Reconstruction results using the **BCS** algorithm. Runtime was $\sim$20 minutes on a standard PC. Shown are the 8 most frequent species. Original strains were : *Escherichia coli, Vibrio fischeri, Staphylococcus epidermidis, Enterococcus faecalis* and *Photobacterium leiognathi* (each with 20% frequency).

Salmonella enterica show a sequence difference in 33 bases over the PCR amplified region, only two bases are different in the region used for the **BCS** reconstruction, and thus the *Escherichia coli* sequence was removed in the database preprocessing stage (see online Supplementary Methods). When this sequence is manually added to the database (in addition to the *Salmonella enterica* sequence), the **BCS** algorithm correctly identifies the presence of *Escherichia coli* rather than *Salmonella enterica* in the mixture. Another strain identified in the reconstruction - the Kennedy Space Center clone KSC6-79 - is highly similar in sequence (differs in five bases over the region tested) to the sequence of *Staphylococcus epidermidis* used in the mixture.

4 Discussion

In this work we have proposed a framework for identifying and quantifying the presence of bacterial species in a given population using information from a single Sanger sequencing reaction. Simulation results with noise levels comparable

to the measured noise in real chromatograms indicate that our method can reconstruct mixtures of tens of species. When not enough information is present in the sequence (for example when a large number of sequences is present in the mixture), performance of the reconstruction algorithm decays gracefully, and still retains detection of the prominent species.

In order to test the applicability of the **BCS** algorithm to real experimental data, we performed a reconstruction of a toy mixture containing five bacterial species. Results of the sample reconstruction (identification of 3 out of 5 species at the strain level, and the additional 2 at the genus level, when E. coli is not omitted from the database) indicate that with appropriate chromatogram preprocessing, **BCS** can be applied to experimental mixtures. However, further optimization of the sequencing and preprocessing is required in order to obtain more accurate results.

Essentially, the amount of information needed for identifying the species present in the mixture is logarithmic in the database size [5, 14], as long as the number of the species present in the mixture is kept constant. Therefore, a single sequencing reaction with hundreds of bases contains in principle a very large amount of information and should suffice for unique reconstruction even when the database contains millions of different sequences. Compressed Sensing enables the use of such information redundancy through the use of linear mixtures of the sample. However, the mixtures need to be RIP in order to enable an optimal extraction of the information. In our case, the mixtures are dictated by the sequences in the database, which are clearly dependent. While two sequences which differ in a few nucleotides have high coherence and clearly do not contribute to RIP, even a single insertion or deletion completely brings the two sequences to being 'out of phase', thus making it easier to distinguish between them using **CS**(provided that the insertion/deletion did not occur to close to the end of the sequenced region). Since the mixing matrix is built using each sequence in the database separately, we do not rely on correct alignment of the database sequences, and, moreover, while a species actually present in the mixture is likely to appear in the solution with high frequency, sequences of similar species which are different by one or a few insertion or deletion events, will violate the linear constrains present in our optimization criteria, and are not likely to 'fool' the reconstruction algorithm.

While limited to the identification of species with known 16S rRNA sequences, the **BCS** approach may enable low cost simple comparative studies of bacterial population composition in a large number of samples. Our method, like any other method, can perform only as well as is allowed by the inherent inter-species variation in the sequenced region. For example, if two species are completely identical at the 16S rRNA locus, no method will be able to distinguish between them based on this locus alone. In the simulations presented, we defined a species reconstruction to be accurate having up to 1 nucleotide difference from the original sequence. Since sequence lengths used were typically around 500bp, the reconstruction sequence accuracy was approx. 0.2%. Average sequence difference between genus was measured as approx. 3%, whereas between species

is approx. 2% [42], and therefore simulation performance was measured at subspecies resolution. However, there are cases of species with identical or nearly identical 16S rRNA sequences, and therefore these species can not be discriminated based on 16S rRNA alone. Sequencing of additional loci (such as in the MLST database [34]) are likely to be required in order to achieve higher reconstruction resolution. Our proposed method can easily be extended to more than one sequencing reaction per mixture, whether they come from the same region or distinct regions, by simply joining all sequencing results as linear constraints. Such an extension can lead to a larger number of linear constraints. This increases the amount of information available for our reconstruction algorithm, which will enable us to both overcome experimental noise present in each sequencing, and distinguish between species more accurately and at a higher resolution.

Acknowledgments

We thank Amit Singer, Yonina Eldar, Gidi Lazovski and Noam Shental for useful discussions, Eytan Domany for critical reading of the manuscript, Joel Stavans for supporting this research and Chaime Priluski for assistance with chromatogram peak prediction data.

References

1. Amann, R., Ludwig, W., Schleifer, K.: Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews 59(1), 143–169 (1995)
2. Armougom, F., Raoult, D.: Use of pyrosequencing and DNA barcodes to monitor variations in firmicutes and bacteroidetes communities in the gut microbiota of obese humans. BMC Genomics 9(1), 576 (2008)
3. Bobin, J., Starck, J., Ottensamer, R.: Compressed sensing in astronomy. Journal of Selected Topics in Signal Processing 2, 718–726 (2008)
4. Brodie, E., DeSantis, T., Parker, J., Zubietta, I., Piceno, Y.M., Andersen, G.L.: Urban aerosols harbor diverse and dynamic bacterial populations. Proceedings of the National Academy of Sciences 104(1), 299–304 (2007)
5. Candes, E.: Compressive sampling. In: Int. Congress of Mathematics, Madrid, Spain, pp. 1433–1452 (2006)
6. Candes, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Arxiv preprint math/0503066 (2005)
7. Candes, E., Tao, T.: Decoding by linear programming. IEEE Transactions on Information Theory 51(12), 4203–4215 (2005)
8. Candes, E., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory 52(12), 5406–5425 (2006)
9. Candes, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics 35(6), 2313–2351 (2007)
10. Dai, W., Sheikh, M., Milenkovic, O., Baraniuk, R.: Compressive sensing dna microarrays. EURASIP Journal on Bioinformatics and Systems Biology (2009), doi:10.1155/2009/162824

11. DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology 72(7), 5069 (2006)
12. Dethlefsen, L., Huse, S., Sogin, M., Relman, D.: The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biology 6(11), e280 (2008)
13. Dewhirst, F., Izard, J., Paster, B., et al.: The human oral microbiome database (2008)
14. Donoho, D.: Compressed sensing. IEEE Transaction on Information Theory 52(4), 1289–1306 (2006)
15. Donoho, D.: For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59(6), 797–829 (2006)
16. Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K., Baraniuk, R.: Single-pixel imaging via compressive sampling. IEEE Signal Processing Magazine 25(2), 83–91 (2008)
17. Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S., Nelson, K., Relman, D.: Diversity of the human intestinal microbial flora. Science 308(5728), 1635–1638 (2005)
18. Erlich, Y., Gordon, A., Brand, M., Hannon, G., Mitra, P.: Compressed Genotyping. IEEE Transactions on Information Theory 56(2), 706–723 (2010)
19. Figueiredo, M., Nowak, R., Wright, S.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE Journal of Selected Topics in Signal Processing 1(4), 586–597 (2007)
20. Gao, Z., Tseng, C., Pei, Z., Blaser, M.: Molecular analysis of human forearm superficial skin bacterial biota. Proceedings of the National Academy of Sciences 104(8), 2927 (2007)
21. Gentry, T., Wickham, G., Schadt, C., He, Z., Zhou, J.: Microarray applications in microbial ecology research. Microbial Ecology 52(2), 159–175 (2006)
22. Guarner, F., Malagelada, J.: Gut flora in health and disease. Lancet 361(9356), 512–519 (2003)
23. Hamady, M., Knight, R.: Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Research 19(7), 1141–1152 (2009), PMID: 19383763
24. Hamady, M., Walker, J., Harris, J., Gold, N., Knight, R.: Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nature Methods 5(3), 235–237 (2008)
25. Hugenholtz, P.: Exploring prokaryotic diversity in the genomic era. Genome Biology 3(2), reviews0003.1–reviews0003.8 (2002)
26. Huse, S., Dethlefsen, L., Huber, J., Welch, D., Relman, D., Sogin, M.: Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genetics 4(11), e1000255 (2008)
27. Kainkaryam, R., Woolf, P.: Pooling in high-throughput drug screening. Current Opinion in Drug Discovery & Development 12(3), 339 (2009)
28. Keller, M., Zengler, K.: Tapping into microbial diversity. Nature Reviews Microbiology 2(2), 141–150 (2004)
29. Kommedal, O., Karlsen, B., Sabo, O.: Analysis of mixed sequencing chromatograms and its application in direct 16S rDNA sequencing of poly-microbial samples. Journal of Clinical Microbiology (2008)

30. Lin, T., Herrmann, F.: Compressed wavefield extrapolation. Geophysics 72 (2007)
31. Lipshutz, R., Taverner, F., Hennessy, K., Hartzell, G., Davis, R.: DNA sequence confidence estimation. Genomics 19(3), 417–424 (1994)
32. Lustig, M., Donoho, D., Pauly, J.: Sparse mri: The application of compressed sensing for rapid MR imaging. Magnetic Resonance in Medicine 58, 1182–1195 (2007)
33. Mager, D., Haffajee, A., Devlin, P., Norris, C., Posner, M., Goodson, J.: The salivary microbiota as a diagnostic indicator of oral cancer: A descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. J. Transl. Med. 3(1), 27 (2005)
34. Maiden, M., Bygraves, J., Feil, E., Morelli, G., Russell, J., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D., et al.: Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences 95(6), 3140–3145 (1998)
35. Medini, D., Serruto, D., Parkhill, J., Relman, D., Donati, C., Moxon, R., Falkow, S., Rappuoli, R.: Microbiology in the post-genomic era. Nat. Rev. Micro. 6(6), 419–430 (2008)
36. Paster, B., Boches, S., Galvin, J., Ericson, R., Lau, C., Levanos, V., Sahasrabudhe, A., Dewhirst, F.: Bacterial diversity in human subgingival plaque. J. of Bacteriology 183(12), 3770–3783 (2001)
37. Savage, D.: Microbial ecology of the gastrointestinal tract. Annual Reviews of Microbiology 31, 107–133 (1977)
38. Sears, C.: A dynamic partnership: Celebrating our gut flora. Anaerobe 11(5), 247–251 (2005)
39. Shental, N., Amir, A., Zuk, O.: Identification of rare alleles and their carriers using compressed se(que)nsing. Nucleic Acid Research 38(19), e179 (2010)
40. Singh, B., Millard, P., Whiteley, A., Murrell, J.: Unravelling rhizosphere-microbial interactions: opportunities and limitations. Trends Microbiol. 12(8), 386–393 (2004)
41. Tropp, J.A.: Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Transactions on Information Theory 52(3), 1030–1051 (2006)
42. Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.H., Ludwig, W., Glckner, F.O., Rossell-Mra, R.: The all-species living tree project: A 16s rrna-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology 31(4), 241–250 (2008)