Introduction

This paper presents a method to generate a >70% accurate prediction of prospectivity of offshore well locations prior to drilling. The approach uses DNA analysis of seabed soil samples to derive information on the mix of microbial species in the samples. Using our database to correlate DNA to soil samples and production data of earlier drilled locations, the new DNA fingerprint is an indicator of the presence of vertical micro-seepage to the surface. This originates from hydrocarbon accumulations in the subsurface - including offshore oil and gas fields. The presence of vertical upward micro-seepage is extensively described in the literature (Laubmeyer, 1933, Horvitz, 1939, Davis, 1956, Sealy, 1974, Miller, 1976, Schumacher, 1996, Wagner, 2002, Schumacher, 2012, Rasheed et al. 2013). It was shown that clear and sharp anomalies of -geochemical- properties could be indicated very precisely at the border of reservoirs due to vertical upward migration through buoyancy (Klusman and Saeed, 1996, Saunders et al. 1999). This causality is used in combination with two relatively new techniques to get more accurate information out of the overly abundant ‘noisy’ signal.

First technological break-through: DNA ‘fingerprinting’, biotechnology: To successfully implement our method it is necessary to determine the complex composition of microbes of the subsurface - not only those that flourish at micro-seepage sites, but also those that are eliminated above hydrocarbon offshore fields. Recent developments in DNA analysis techniques have made this complex and previously expensive problem efficiently and economically solvable.

Second technological break-through: Big Data, Machine Learning, supercomputing: resulting measurement data leads to terabytes of data which must be correlated to hydrocarbon presence. Advancements in machine learning applications together with parallel computing (Hadoop in the cloud, using GPU’s) have made it possible to construct robust and reliable predictive DNA-based models for hydrocarbon accumulations in the subsurface. In this study we illustrate the extended technology with an offshore case study: a ranking of prospects in the North Sea.

Method

To generate an accurate prediction of the prospectivity of a location, the following four steps are applied (for a more extensive description see Te Stroet et al. 2017):

1. Soil sampling in the field, followed by DNA analysis to obtain DNA-fingerprints of microbes.

The methodology described here is derived from the workflow to determine a DNA-fingerprint is in life sciences applications (extraction, multiplication with PCR and sequencing of DNA), modified for our purpose. Figure 1 shows a schematic representation of the DNA-based analysis of soil samples.

Figure 1 Workflow for obtaining DNA fingerprints for soil samples: 1) Soil samples preparation for DNA analysis, 2) Extraction of bacterial DNA from soil sample, 3) Amplification of the 16S rDNA by Polymerase Chain Reaction, 4) 16S Sequence analysis using next generation Illumina MiSeq, 5) Processing raw data from MiSeq to verified 16S sequences, 6) Processing verified 16S sequences back to individual soil samples and 7) Interpretation of 16S sequence data: translation to bacterial genera (families); steps 2 and 3 are tuned by Biodentify to get maximum information on species.
2. Selecting a training set from our database from earlier drilled areas with known prospectivity. The training set has similar DNA fingerprints compared to the new samples and can correlate DNA from soil samples with oil or gas production data. By selecting the samples from the database that have highest correlation with DNA fingerprints of the new samples a training set is generated, where fingerprints and productivity are known. This training set is used to find the correlation model with Machine Learning algorithms to predict the prospectivity at new locations (see step 3).

3. Modelling and validating the microbes that determine prospectivity. Results are correlated to the sequenced analysis (the bacterial diversity in the DNA-fingerprints of all soil samples) with production data in the selected training set. The goal is to find correlations between production and presence or absence of specific bacteria, typically about 50-200 out of the 340,000 species in our database. These are called biomarkers. At least 50 biomarkers are needed to accurately estimate the location of hydrocarbon accumulation (Figure 2c). A model is built with methods that are used to deal with 'sparse modelling' issues like non-linearity, influence of noise and prevention of overfitting (Gaussian kernels for non-linear problems (Cortes, 2012) and L1-based regularization methods (Mosci, 2011)).

![Figure 2a) area with hydrocarbons (green) and no hydrocarbons (red) b) modelled prospectivity using 1 biomarker c) modelled prospectivity using 70 biomarkers.](image)

4. Analysing prospectivity of offshore prospects. Next DNA-fingerprints are correlated with prospectivity. This can pose a problem in the current project as not many seabed samples are available and they are expensive to obtain. An innovative step is used to predicting seabed samples: a correlation model made using stored cuttings from drilled wells. This is possible because the microbial ecosystem does not evolve anymore when pressure and temperature conditions are changed abruptly; the DNA-fingerprint ‘freezes’ and DNA is stored forever.

**Example**

The four steps are applied in this project as follows:

1. Determining a representative data set of about 1000 locations; about 500 producing locations and 500 dry locations, see Figure 3 (data through Dutch database NLOG www.nlog.nl/).
2. Analysing DNA fingerprints of the cuttings
3. Building predictive model that differentiates producing and non-producing sampled locations
4. Predict prospectivity at seabed sampled locations made available by two North Sea operators.

![Figure 3 Example of sampled cuttings in the Dutch core shed (left), and locations of cuttings sampled to build the training model (right).](image)
Steps 1-3: Using Cuttings
A model is ‘trained’ on a different random 70% subset of the training data and validated on 30% of the samples that are left out from the training set (see Figure 4). Each location receives a predicted value between -1 and 1 whether the sample is above a hydrocarbon field. This is repeated 1000 times with different random sets (see Figure 5) leading to all locations being used on average 300 times to validate the models and accurate statistics of prediction accuracy.

![Figure 4](image1.png) 70% of the training set is used for building one predictive model (left); this model is validated on 30% of the randomly left out part of the training set (right)

![Figure 5](image2.png) The process described in figure 4 is repeated 1000 times and thus there are 1000 validated models. Resulting prediction accuracy >80% (827 non-white points were predicted correctly).

De-risking drilling of prospects: When the prediction accuracy of the previous steps is satisfactory, the trained models can be used with DNA fingerprints of new locations (where presence of hydrocarbons in the subsurface is unknown) as input. In other words a prospectivity indicator is created, where new samples can be given a value -1 (no hydrocarbons expected present) to 1 (hydrocarbons expected to be present). The project is currently in the stage of generating prospectivity estimates: two oil and gas operators delivered seabed/vibrocore samples. The seabed samples are above new North Sea prospects. Prospectivity was already estimated based on earlier training with only 48 of the 250 locations above known oil fields/dry structures (i.e. known prospectivity). Training with the 1000 cuttings aims to increase the accuracy significantly.
Conclusions

This study shows that cuttings can be used to train a reliable predictive model for estimating prospectivity for offshore hydrocarbon targets. In this example a predictive accuracy of 80% was obtained. Furthermore analysing seabed samples taken around target well locations or across new prospects, and applying DNA analysis and machine learning on these samples can significantly reduce drilling and development risk by indicating the relative prospectivity of the planned wells or prospects.

Acknowledgements

The authors are acknowledging Dana Petroleum, Dutch Topsector Gas, EBN, SME Grant of European Commission and Wintershall for contributing to this research.

References


